Università di Catania

ASSOCIAZIONE per l'INFORMATICA UMANISTICA e la CULTURA DIGITALE

Consiglio Nazionale delle Ricerche

ME.TE. DIGITALI

MEDITERRANEO IN RETE TRA TESTI E CONTESTI

ATTI DEL XIII CONVEGNO ANNUALE
AIUCD 2024

28 - 30 MAGGIO

MONASTERO DEI BENEDETTINI

P.ZZA DANTE, 32 CATANIA

ISBN 978-88-942535-8-0

A cura di: Di Silvestro Antonio; Spampinato Daria (2024). Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD, Catania 28-30 maggio 2024, Università di Catania.

Editing: Denise Bruno; Christian D'Agata; Laura Mazzagufo; Francesca Prado; Eliana Vitale; Alessandro Zammataro.

Ultimo accesso agli URL in data 15 maggio 2024.

Si prega di notificare all'editore ogni omissione o errore si riscontri: aiucd.segreteria [at] aiucd.org
Please notify the publisher of any omissions or errors found: aiucd.segreteria [at] aiucd.org

Il programma della conferenza AIUCD 2024 è disponibile online https://aiucd2024.unict.it/programma/
The AIUCD 2024 Conference Program is available online https://aiucd2024.unict.it/programma/

I contributi pubblicati nel presente volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima mediante *double-blind peer review* sotto la responsabilità del Comitato di Programma di AIUCD 2024.
All the papers published in this volume have received favourable reviews by experts in the field of DH, through an anonymous double-blind peer review process under the responsibility of the AIUCD 2024 Program Committee.

Chair

Antonio Di Silvestro (Università di Catania)
Daria Spampinato (CNR Istituto di Scienze e Tecnologie della Cognizione)

Comitato di programma / *Program committee*

Emmanuela Carbé (Università di Siena)
Massimo Cultraro (CNR Istituto di Scienze del Patrimonio Culturale)
Christian D'Agata (Università di Catania)
Antonio Di Silvestro (Università di Catania)
Greta Franzini (Eurac Research)
Maurizio Lana (Università del Piemonte Orientale)
Cristina Marras (CNR Istituto del Lessico intellettuale europeo e Storia delle Idee)
Marco Mazzone (Università di Catania)
Ouafae Nahli (CNR Istituto di Linguistica Computazionale "Antonio Zampolli")
Marianna Nicolosi-Asmundo (Università di Catania)
Marina Paino (Università di Catania)
Giuseppe Palazzolo (Università di Catania)
Jonathan Prag (University of Oxford Merton College)
Daria Spampinato (CNR Istituto di Scienze e Tecnologie della Cognizione)
Rachele Sprugnoli (Università di Parma)
Francesco Stella (Università di Siena)

Segreteria scientifica / *Scientific Secretariat*

Liborio Barbarino (Università di Catania)
Denise Bruno (Università di Catania)
Giulia Cacciatore (Università di Catania)
Giuseppe Canzoneri (Università di Catania)
Elisa Conti (Università di Catania)
Milena Giuffrida (Università di Catania)
Miryam Grasso (Università di Catania)
Francesca Prado (Università di Catania)
Emilio M. Sanfilippo (CNR Istituto di Scienze e Tecnologie della Cognizione)
Eliana Vitale (Università di Catania)
Alessandro Zammataro (Università di Catania)

Comunicazione istituzionale: Claudia Cantale (Università di Catania) e Area Per la Comunicazione dell'Università di Catania (ACOM).
Institutional communication: Claudia Cantale (University of Catania) and the Area for Communication of the University of Catania (ACOM)

Supporto tecnico: Rosario Agrò, Area della Terza Missione dell'Università di Catania, per la consulenza e la progettazione grafica dei materiali informativi del convegno.
Technical support: Rosario Agrò, Third Mission Area of the University of Catania, for advice and graphic design of the conference information materials.

Enti organizzatori / *Organisers*
AIUCD; Università di Catania: Dipartimento di Scienze Umanistiche; CNR Istituto di Scienze e Tecnologie della Cognizione; CINUM: Centro di Informatica Umanistica dell'Università di Catania.

Supporter
CLARIN-IT; Neperia Group; Storage; programma Piaceri 2020-2022, Linea 1; Parmalat-Sole.

Chair di area/ Track chair

Le culture digitali nel Mediterraneo
Cristina Marras (CNR Istituto del Lessico intellettuale europeo e Storia delle Idee)
Paola Moscati (CNR Istituto di Scienze del Patrimonio Culturale)
Archivi ed edizioni digitali
Christian D'Agata (Università di Catania)
Greta Franzini (Eurac Research)
Analisi computazionale dei testi
Angelo Mario Del Grosso (CNR Istituto di Linguistica Computazionale "Antonio Zampolli")
Simone Rebora (Università di Verona)
Ontologie e Semantic Web
Marianna Nicolosi Asmundo (Università di Catania)
Francesca Tomasi (Università di Bologna)
Preservazione della memoria e del patrimonio digitale
Fabio Ciracì (Università del Salento)
Anna Maria Marras (Università di Torino)

Lista dei revisori /List of reviewers

Maristella Agosti (Università di Padova), **Stefano Allegrezza** (Università di Bologna), **Chiara Alzetta** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Liborio Barbarino** (Università di Catania), **Nicola Barbuti** (Università di Bari Aldo Moro), **Stefano Bazzaco** (Università di Verona), **Benedetta Bessi** (Università Ca' Foscari di Venezia), **Andrea Bolioli** (ricercatore indipendente), **Paolo Bonora** (Università di Bologna), **Federico Boschetti** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Dominique Brunato** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Flavia Bruni** (Università Gabriele D'Annunzio di Chieti-Pescara), **Marina Buzzoni** (Università Ca' Foscari di Venezia), **Alberto Campagnolo** (Universitè Catholique de Louvain/KULeuven), **Anna Cappellotto** (Università di Verona), **Emmanuela Carbé** (Università di Siena), **Vittore Casarosa** (CNR Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – Università di Pisa), **Fabio Ciotti** (Università di Roma "Tor Vergata"), **Fabio Ciracì** (Università del Salento), **Elisa Conti** (Università di Catania), **Salvatore Cristofaro** (CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee), **Christian D'Agata** (Università di Catania), **Elisa D'Argenio** (HUN-REN Hungarian Research Centre for Linguistics), **Mauro De Bari** (Università di Bari Aldo Moro), **Riccardo Del Gratta** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Angelo Mario Del Grosso** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Saulo Delle Donne** (Università del Salento), **Giorgio Maria Di Nunzio** (Università di Padova), **Antonio Di Silvestro** (Università di Catania), **Filippo Diara** (Università di Torino), **Giulia Fabbris** (Università Ca' Foscari di Venezia), **Riccardo Fedriga** (Università di Bologna), **Franz Fischer** (Università Ca' Foscari di Venezia), **Greta Franzini** (Eurac Research), **Francesca Frontini** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Daniele Fusi** (Stuttgart University & VeDPH – Università Ca' Foscari di Venezia), **Carola Gatto** (Università del Salento), **Lucia Giagnolini** (Università di Bologna), **Emiliano Giovannetti** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Milena Giuffrida** (Università di Catania), **Edmondo Grassi** (Università San Raffaele di Roma), **Miryam Grasso** (Università di Catania), **Alessandro Iannella** (Università di Cagliari - Università di Pisa – Università di Torino), **Paola Italia** (Università di Bologna), **Maurizio Lana** (Università del Piemonte Orientale), **Pietro Maria Liuzzo** (Bibliotheca Hertziana), **Dominique Longrée** (Université de Liège), **Francesco Mambrini** (Università Cattolica del Sacro Cuore di Milano), **Tiziana Mancinelli** (Istituto Italiano di Studi Germanici), **Anna Maria Marras** (Università di Torino), **Cristina Marras** (CNR Istituto del Lessico intellettuale europeo e Storia delle Idee), **Federico Meschini** (Università della Tuscia), **Alessio Miaschi** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Paolo Monella** (Università Sapienza di Roma), **Ouafae Nahli** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Marianna Nicolosi-Asmundo** (Università di Catania), **Giuseppe Palazzolo** (Università di Catania), **Valentina Pasqual** (Università di Bologna), **Gianluca Pavani** (Università di Roma "Tor Vergata"), **Giulia Pedonese** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Jonathan Prag** (University of Oxford Merton College), **Simone Rebora** (Università di Verona), **Giulia Renda** (Università di Bologna), **Roberto Rosselli Del Turco** (Università di Torino), **Enrica Salvatori** (Università di Pisa), **Emilio M. Sanfilippo** (CNR Istituto di Scienze e Tecnologie della Cognizione), **Eva Sassolini** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Pietro Sichera** (CNR Istituto di Scienze e Tecnologie della Cognizione), **Daniele Silvi** (Università di Roma "Tor Vergata"), **Elena Spadini** (University of Basel), **Daria Spampinato** (CNR Istituto di Scienze e Tecnologie della Cognizione), **Linda Spinazzè** (Università Ca' Foscari di Venezia), **Rachele Sprugnoli** (Università di Parma), **Francesco Stella** (Università di Siena), **Cecilia Tamagnini** (Università di Bologna), **Timothy Tambassi** (Università Ca' Foscari di Venezia), **Francesca Tomasi** (Università di Bologna), **Marco Venuti** (Università di Catania), **Fabio Vitali** (Università di Bologna).

SOMMARIO

Prefazione

Il XIII convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale ha nel titolo un acronimo che è anche una parola semanticamente piena: *Me.Te. [Digitali]*. L'intento è quello di mettere su un piano di 'parità' da un lato la dimensione geopolitica delle Digital Humanities, nello spazio dialogico per eccellenza della cultura europea, oggi drammatico teatro di conflitti che si acuiscono sempre di più, e dall'altro le diverse forme della testualità, intesa nella sua accezione etimologica di intreccio di parole, immagini, suoni, regolato da forme di coesione e coerenza, e declinato in una forma il più possibile inclusiva, in quanto veicolo primario dell'organizzazione della conoscenza. L'aggettivo *digitale* può essere riferito sia alla seconda parte dell'acronimo (i testi, come chiarito nel sottotitolo - *Mediterraneo in rete tra testi e contesti* -, sono anche *con-testi*), sia all'acronimo nella sua interezza: le *mete digitali* rappresentano per noi il senso di un processo volto a creare connessioni tra testi e persone, istituire dialoghi a distanza tra culture diverse, realizzare spazi virtuali di condivisione di testi e artefatti riconducibili a una radice culturale condivisa.
Per la prima volta il convegno AIUCD ha luogo in Sicilia, spazio elettivo (saremmo tentati di dire simbolico) del Mediterraneo. Non è nostro intento porre la questione geopolitica del rapporto centro-periferia, che nella prospettiva del ruolo delle DH nel Sud del mondo potrebbe non fornire delle risposte operative adeguate. Si tratta piuttosto per un verso di guardare in una prospettiva glo-cale, per l'altro di ridare credito, pur nella condivisione di metodologie e pratiche, alle peculiarità di comunità di ricerca identificate da specifiche tradizioni di studi e radici culturali. Condivisibile quanto scrive Fiormonte: "Le Digital Humanities del Sud oggi hanno l'opportunità non tanto di sostituirsi o sovrapporsi alle realtà ancora dominanti, ma di diventare il punto di riferimento di modelli etimologicamente plurali e sostenibili di conservazione, accesso e trasmissione della conoscenza in formato digitale".[1]

***

Il Convegno AIUCD2024 vede l'organizzazione congiunta del Dipartimento di Scienze Umanistiche dell'Università di Catania e dell'Istituto di Scienze e Tecnologie della Cognizione del Consiglio Nazionale delle Ricerche di Catania. Alla base di questa collaborazione vi è un forte dialogo e una consolidata collaborazione scientifica tra diversi studiosi del DiSUm e ricercatori dell'ISTC, collaborazione di cui i due chair del convegno, nella qualità rispettivamente di Direttore del Centro di Informatica Umanistica (CInUm), centro interdipartimentale attivato in seno al DiSUm, e di responsabile della sede locale dell'ISTC-CNR, fungono da tramiti e 'facilitatori'. È una collaborazione, tra l'altro, tra un Dipartimento universitario e un Istituto appartenente a un Ente di ricerca, tra umanisti e informatici che da anni operano nel settore della lessicografia e filologia digitale e dell'umanistica computazionale, che hanno trovato nell'AIUCD una casa comune e una comunità con cui confrontarsi e in cui crescere.
Un riferimento alla storia e all'attività dei due enti che hanno collaborato ci pare necessario. Il Dipartimento di Scienze Umanistiche, nella sua primitiva *facies* di Facoltà di Lettere e Filosofia, vede nascere negli anni '80, grazie all'intuizione e agli studi di Giuseppe Savoca, una tradizione di lessicografia applicata ai testi letterari italiani dell'Otto/Novecento. Già in quegli anni Savoca portava avanti le proprie ricerche attraverso una rete nazionale, costituita dal CLIPON (Gruppo Nazionale di Coordinamento del Consiglio Nazionale delle Ricerche per le "Concordanze della Lingua Italiana Poetica dell'Otto/Novecento", di cui Savoca è stato promotore e presidente del Consiglio scientifico), che negli anni 1984-2000 ha visto riunite le Università di Catania, Roma Sapienza, Roma Tor Vergata, Torino, Cagliari, Urbino e Lecce. Ma altrettanto fervido di risultati ed esperienze era il gruppo di lavoro 'locale', costituito dal Dottorato di ricerca in "Italianistica – Lessicografia e Semantica del Linguaggio Letterario Europeo", che nel corso della sua quasi venticinquennale attività (dal 1989 al 2012), ha prodotto in collaborazione numerosi lavori di concordanza pubblicati nella prestigiosa collana degli "Strumenti di Lessicografia Letteraria Italiana" dell'editore Olschki. Uscendo dalle secche di un'informatica puramente linguistica, Savoca ha puntato i suoi sforzi su una vocabolarizzazione delle grandi opere in versi italiane del XX secolo (realizzando per Zanichelli a metà degli anni '90 un 'inedito' *Vocabolario della poesia italiana del Novecento*), in cui il rigore della classificazione propria di una concordanza avesse alla propria base una visione del segno linguistico 'triplice' basata su significante, significato e frequenza. Questa visione 'inclusiva' della parola letteraria ha fatto sì che la lessicografia si aprisse verso le ragioni della filologia e dell'interpretazione puntuale, nella prospettiva che un vocabolario-concordanza, puntando verso la semantica profonda dei testi, è un'operazione di vera e propria ri-testualizzazione. Per queste ragioni storiche, metodologiche e scientifiche, abbiamo voluto che ad aprire i lavori del convegno come keynote

[1] D. Fiormonte, G. Del Rio Riande, *Dalla periferia all'impero. Digital Humanities e diversità culturale*, in *Digital Humanities. Metodi, strumenti, saperi*, a cura di F. Ciotti, Roma, Carocci, 2023, p. 370.

fosse proprio Savoca, il cui intervento porta come titolo “Informatica umanistica, Infocrazia, automi e intelligenze artificiali”.

Il Centro Informatica Umanistica, attivato nel 2017, nasce dalla volontà di proseguire la tradizione di lessicografia computazionale aperta da Savoca, collegandosi a una metodologia volta a coniugare le ragioni della lessicografia, della filologia e dell’interpretazione (*Lessicografia, filologia e critica* è il titolo del convegno ‘inaugurale’ tenutosi a Catania a metà degli anni ’80). Ricollegandosi al “Centro di informatica letteraria” fondato da Savoca, il CInUm ha puntato soprattutto su una tradizione filologico-linguistica che ha formato nuove generazioni di studiosi (diversi dei quali oggi ricercatori anche su fondi PNRR), con i quali sono state messe a punto modalità innovative, multi-metodologiche e *lexicon-based* di edizioni critiche digitali dei testi letterari italiani moderni e contemporanei. Da questo sforzo nasce l’impresa – voluta dalla Commissione Nazionale dell’Opera Omnia nominata dal MiBACT e posta sotto l’egida di Mondadori – dell’Edizione Digitale dell’Opera Omnia di Pirandello (https://www.pirandellonazionale.it), complementare all’Edizione nazionale cartacea, e interamente gestita dal CInUm con la direzione di Antonio Sichera, membro della Commissione, e dallo scrivente Unict di questa introduzione, direttore del CInUm. Il progetto aggiorna ai nuovi orizzonti metodologici delle *Digital Scholarly Editions* e apre a diversi livelli di utenza (dallo specialista al lettore ‘ingenuo’) quei livelli di analisi linguistica e filologica dei testi già messi a punto da Savoca. Nel tempo-struttura dell’edizione, grazie all’apporto di giovani leve di umanisti digitali, il progetto è passato dalla forma dell’edizione-archivio, già di per sé innovativa per la presenza di una zona di strumenti didattici e percorsi di senso realizzati con tecnologie di analisi e interrogazione dei testi, a quella di una hyperedizione che rappresenta essa stessa un’idea di rete, ma non nel senso ‘vulgato’ dell’ipertesto, bensì in quello del dialogo tra metodologia filologica, lessicografica, ermeneutica e didattica. Tale interconnessione, resa possibile dall’intreccio tra codifica in XML/TEI, vocabolari lemmatizzati dei testi, riflessione linguistica ed ermeneutica e risorse multimediali, fa dell’Edizione Digitale di Pirandello un esempio originale nel panorama internazionale, ma soprattutto un grande laboratorio ‘reticolare’ di esperienze, competenze, sperimentazioni, che si arricchiscono grazie all’entusiasmo di giovani continuamente formati all’interno del CInUm. Un percorso che, grazie al sostegno dello storico editore di Pirandello, ha dato i suoi frutti, sia per il senso del lavoro di equipe sia per l’utilizzo funzionale e ‘dinamico’ della filologia per il commento ai testi, anche nella poderosa edizione cartacea dell’*Opera poetica* di Pavese, curata sempre dai responsabili dell’edizione digitale di Pirandello (ma con il concorso di numerosi curatori delle varie sezioni). Questo percorso, che è stato un lungo e costante itinerario di formazione dei ricercatori in progetti tra cartaceo e digitale, ha fatto sì che alcuni di essi stiano lavorando su un importante progetto di edizione digitale critica e commentata dei testi dei grandi veristi ottocenteschi (“Verismo Digitale”), di cui è PI la Direttrice del DiSUm Marina Paino, che condivide la responsabilità scientifica con Andrea Manganaro e Antonio Sichera, e si avvale dell’expertise del Direttore del CInUm e di Giuseppe Palazzolo. Questo progetto, che rappresenta la costola catanese dello Spoke 3 PNRR “Digital Libraries, Archives and Philology” nell’ambito del Partenariato Esteso CHANGES (Cultural Heritage Active Innovation for Sustainable Society), verrà presentato all’interno del panel di “Filologia digitale”.

Le DH rappresentano una delle linee strategiche della ricerca del DiSUm, e hanno anche una forte dimensione didattica. Infatti il DiSUm ha attivato da qualche anno, con ottimi riscontri e risultati lusinghieri, grazie a una forte convergenza di studiosi di diverse discipline e alla grintosa *leadership* del collega Marco Mazzone, un corso di laurea magistrale in “Scienze del testo per le professioni digitali”, che punta su una formazione *text-based* e che si sta sempre più consolidando attraverso un dialogo virtuoso tra didattica, ricerca e mondo aziendale (di cui Neperia Group è il rappresentante ‘storico’ extrauniversitario del corso).

Veniamo adesso all’altro ‘partner. L’Istituto CNR di Scienze e Tecnologie della Cognizione è un Istituto interdisciplinare che, anche se afferisce a un Dipartimento Scienze Umane e Sociali, Patrimonio Culturale, porta avanti diverse attività di ricerca che confluiscono nel Dipartimento di ICT, e ha sviluppato proprio nella sede di Catania temi di ricerca specifici e caratterizzanti. Temi che spaziano tra intelligenza artificiale e sistemi tecno-sociali (robotica, *decision making*, tecnologie semantiche, comprensione automatica delle lingue naturali), neuroscienze computazionali, processi cognitivi, comunicativi e linguistici (teoria e analisi del parlato e della variabilità linguistica, dialetti, prosodia, azione gesto e lingue dei segni), sviluppo cognitivo, apprendimento e socializzazione nei bambini e nei primati non umani, qualità dell'ambiente, salute e società.

In particolare, a Catania opera il gruppo di ricerca CHROMA (*Computational Humanities: Representation, Organization, Management, and Analysis*), che si occupa di rappresentazione, organizzazione, gestione e analisi di diverse tipologie di dati e metadati attraverso metodi computazionali. I domini principali di interesse sono l'informatica umanistica e il patrimonio culturale; le ricerche sono caratterizzate da una connotazione fortemente interdisciplinare, e prevedono lo sviluppo e il riutilizzo di buone prassi e di strumenti *open source standard* per i domini di riferimento. Il gruppo è composto principalmente da informatici, ma collabora costantemente con umanisti di diversi domini, dall’epigrafia e storia romana alla musicologia, dall'archivistica alla letteratura e filologia italiana. La competenza nella modellazione dei dati e nel

trattamento computazionale del testo si è indirizzata sia verso il patrimonio epigrafico del Museo civico di Catania, col progetto EpiCUM (http://epicum.istc.cnr.it), con una importante componente di utilizzo delle ontologie per i beni culturali, sia verso i progetti multidisciplinari di edizione scientifica digitale, tra cui importante è quella delle lettere autografe di Vincenzo Bellini (Bellini Digital Correspondence http://bellinicorrespondence.cnr.it), edita dal CNR con un proprio ISBN (vantato da pochissime edizioni digitali). Il modello di lavoro utilizzato, distribuito, collaborativo e cooperativo, con una forte connotazione didattica universitaria, non è circoscritto al dominio di riferimento ed è, quindi, riusabile in contesti simili.
La collaborazione tra ISTC-CNR e DiSUm, oltre che con il Centro di Informatica Umanistica, collaborazione che ha portato anche all'attivazione di tirocini comuni per studenti e alla presentazione di progetti PRIN (di cui due da poco approvati e attivi), concerne anche lo studio dei processi linguistici e la lingua dei segni, le tematiche di Semantic Web e Digital Humanities.
Questa connotazione fortemente interdisciplinare e dialogica delle attività dell'ISTC-CNR ci ha indotto a puntare, come seconda keynote, su Tara Andrews, la cui storia professionale merita anch'essa una menzione. Infatti (cosa pressoché impossibile per lo stato dei comparti disciplinari in Italia) Andrews non solo è professoressa di Digital Humanities presso l'Istituto di Storia dell'Università di Vienna, ma proviene da una formazione universitaria in Scienze umanistiche e Ingegneria presso il Massachusetts Institute of Technology. Una doppia formazione scientifica e un'esperienza professionale nell'industria del software che ne hanno arricchito il profilo di prospettive originali e innovative sull'uso dei metodi digitali e computazionali nei settori umanistici. La lezione di Andrews si intitola "The medieval Mediterranean in...data? Interpretation, conjecture and digital methods"; essa verterà sui modi in cui i metodi digitali intervengono nello studio della storia – in particolare delle società medievali del Mediterraneo –, e porrà l'attenzione su come affrontare la sfida più ampia di analisi computazionali eseguite su insiemi di dati che saranno sempre soggetti a congetture e persino a controversie.

***

Veniamo adesso alle risposte ricevute alla call for papers. Come nei precedenti convegni, e come emerge dall'organizzazione delle sessioni di lavoro, sono stati individuati degli assi metodologici, volutamente trasversali ai comparti disciplinari, uno dei quali specificamente declinato sul tema del convegno: 1. Le culture digitali nel Mediterraneo; 2. Archivi ed edizioni digitali; 3. Analisi computazionale dei testi; 4. Ontologie e Semantic Web; 5. Preservazione della memoria e del patrimonio digitale. Abbiamo voluto raggruppare gli articoli dei *proceedings* secondo una ulteriore partizione per dare maggiore contezza degli assi di interesse, e quindi abbiamo tracciato 8 percorsi tematici: Mediterraneo tra testi e contesti; Archivi e musei digitali per il patrimonio culturale; Edizioni Scientifiche Digitali; Dizionari e digitalizzazione di banche dati; Analisi computazionale, Intelligenza Artificiale e linguistica; Organizzazione della conoscenza con Semantic Web; Public History e archeologia digitale; Infrastrutture per la ricerca umanistica.
Proprio per la natura multidisciplinare di AIUCD, alle cui conferenze annuali si incontrano studiosi delle diverse discipline che fanno capo alle cosiddette *Digital Humanities* (linguisti, storici, filosofi, archeologi, archivisti, filologi, umanisti digitali, informatici, ingegneri informatici e non solo), sono state presentate proposte supportate da metodologie robuste, solida bibliografia, riutilizzo di strumenti, prassi e standard consolidati e metodi innovativi in linea con i principi FAIR (*Findability, Accessibility, Interoperability, Reusability*) per i dati e TRUST (*Transparency, Responsibility, User focus, Sustainability, Technology*) per le infrastrutture. Tuttavia, anche a causa del 'volano' legato alle richieste del versante digitale del PNRR, molti progetti che prevedono l'uso del digitale sono ancora in una fase iniziale, e quindi non sempre sono supportati da una piena consapevolezza delle metodologie e degli strumenti utilizzabili.
Veniamo ai 'numeri' del convegno. Sono pervenute 117 proposte, di cui 16 non hanno superato una valutazione positiva, mentre una proposta è stata ritirata. La prima area ha ricevuto una discreta attenzione, con contributi che toccano anche l'altra sponda del Mediterraneo, anche se una quota maggioritaria delle proposte afferisce, in continuità con i convegni precedenti, alla seconda area tematica.
Va anche sottolineato che la componente 'plurilingue' dei relatori presenti al convegno, non molto ampia, è stata compensata dai seminari e workshop tenuti nell'ambito di "Aspettando AIUCD2024", cui hanno partecipato studiosi greci, croati, marocchini, trattando tematiche che abbracciano anche i corpora dei dialetti arabi. Un grazie sentito pertanto a tutti i colleghi che hanno voluto contribuire ad arricchire il Convegno con le loro lezioni. Da metà marzo fino a fine maggio con cadenza bisettimanale si sono svolti i seguenti seminari: Federico Boschetti (CNR Istituto di Linguistica Computazionale, IT), "Acquisizione del testo digitale dalle immagini: OCR e HTR"; Cristina Marras (CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, IT), "Migrazioni di tecnologie e linguaggi. Teorie e pratiche"; Neven Jovanović (Sveučilište u Zagrebu, Filozofski fakultet, HR), "Wikidata as infrastructure for an analysis of Latin architectural

terminology"; John Pavlopoulos (Athens University of Economics and Business, Department of Informatics, EL), "Machine Learning for Ancient Languages"; Daniel Riaño Rufilanchas (CSIC-Instituto de Lenguas y Culturas del Mediterráneo y Oriente Próximo, ES), "Collaboration and Innovation in Digital Humanities: The Papyri.info Ecosystem"; Nadia Khlif (Mohammed First University in Oujda, Morocco - CNR Istituto di Linguistica Computazionale, IT), "Challenges and Progress in Constructing Arabic Dialect Corpora and Linguistic tools to Arabic Moroccan Dialect"; Paolo Monella (Università Sapienza, Roma, IT), "Wikipedia e Wikibooks per l'insegnamento delle materie letterarie". Il ciclo era fruibile attraverso una piattaforma digitale da remoto, e questo ha consentito una straordinaria risposta all'iniziativa: si sono avute 241 adesioni all'evento, con diversi non italiani, una cinquantina di studiosi provenienti dal mondo della scuola, delle biblioteche pubbliche, quindi esterni al mondo dell'accademia, ma dotati di grande sensibilità verso un orizzonte fondamentale per il futuro delle DH: quello della costruzione di percorsi di formazione sempre più funzionali per coloro che lavorano essi stessi come formatori nell'ambito del digitale.

***

Va da sé che nostra volontà era quella di essere il più possibile inclusivi, facendo sì che multidisciplinarietà fosse in primis multiculturalità; ma eravamo altrettanto consapevoli che le reti (termine forse ormai *demodé*) fossero in primo luogo quelle digitali, da declinare con sempre maggiore convinzione nella direzione dell'accessibilità, facendo in modo (ci piace dirlo con le parole introduttive al convegno AIUCD del 2021) che "le persone con disabilità sensoriali, motorie o cognitive, o i discenti nelle varie fasi dell'età evolutiva, non *siano* affatto beneficiari passivi di risorse e strumenti creati per loro, ma sono protagonisti attivi nella progettazione di risorse, strumenti, percorsi ideati e, grazie a questo, migliori per tutti".
Per fare questo, occorre operare nella convinzione profonda che la pratica del digitale non può più essere assoggettata a una visione 'strumentalista', per la quale "i metodi e le infrastrutture digitali modificano le condizioni di possibilità della ricerca, l'accesso alle fonti (attraverso la digitalizzazione di archivi e biblioteche) o le modalità di produzione e disseminazione dei risultati, ma non hanno necessariamente un impatto trasformativo su ciò che definisce la ricerca umanistica".[2] Quest'ultima non può più rinchiudersi nella *turris eburnea* della 'preziosità' dei suoi risultati prodotti 'manualmente', riservandosi l'esclusività di legiferare sul 'cosa' e delegando acriticamente il 'come' al lavoro dell'informatica. Occorre insistere su un paradigma 'reticolare', di mutua compenetrazione e vicendevole trasformazione, per il quale il rapporto con l'informatica consente l'elaborazione di una metodologia che incide sul modo di fare, produrre, disseminare la ricerca umanistica. Solo così, ci piace parafrasare sul titolo e sottotitolo del convegno, le reti possono diventare mete digitali inclusive, multidisciplinari, proiettate verso il futuro.
Riflessione su metodologie, e quindi su modellizzazione e linguaggi, sono a nostro avviso il salutare antidoto per contravvenire per un verso all'asservimento alle tecnologie 'imposte' dalle grandi multinazionali dell'High Tech, per l'altro, fatto ancora più importante, il modo migliore per gestire con saggezza (atteso che l'informatica umanistica non si occupa tendenzialmente di *Big Data*) quell'inevitabile processo di "politicizzazione" del testo che è "il frutto della digitalizzazione di ogni aspetto, di ogni frammento, di ogni interstizio della nostra vita".[3] La datizzazione acritica e massiva va sempre gestita, anzi convogliata verso uno sforzo di interpretazione dei fenomeni, ed è all'istanza dell'interpretazione che la ricerca dell'informatica umanistica deve sempre mirare, ricordando (vichianamente) che *verum et factum convertuntur.*
In questa fase più matura delle DH si tratta di capire non solo come utilizzare le tecnologie, diffonderle, renderle patrimonio della comunità italiana in primo luogo, ma anche prevedere le derive e le storture che il digitale, il monopolio, la gestione privacy possono portare. Bisogna acquisire consapevolezza, ragionare e fornire modelli interpretativi e buone prassi. Questa è un'altra grande sfida che dobbiamo cogliere, e con questo convegno ci auguriamo di aver tracciato un primo solco che non si spera non passi inosservato.
Il convegno si arricchisce a partire dai contributi pervenuti e che sono stati valutati come adeguati al termine del processo di referaggio, ed è un'occasione perché la comunità AIUCD si confronti e crei nuove opportunità di collaborazioni scientifiche. Insieme a questo, la presenza dei *keynote* è portatrice di esperienze e riflessioni di amplissimo respiro, e fornisce un contributo fondamentale alla rilevanza scientifica dell'incontro, ponendo domande e interrogativi da cui AIUCD trarrà materia di riflessione e spunti fruttuosi per i prossimi convegni.
Abbiamo voluto anche dare un risalto maggiore ai temi portanti del Convegno organizzando due sessioni dedicate alle *Me.Te. Digitali*. La prima, dal titolo "Mediterraneo tra testi e contesti", è aperta da Salvatore Capasso con una relazione dal titolo "Il Mediterraneo: un mare di opportunità e sfide", in cui lo studioso analizza le sfide geopolitiche ed economiche globali nello scenario chiave del Mediterraneo, con particolare attenzione alle sfide che la transizione digitale e il connesso

[2] F. Ciotti, *Introduzione. La galassia delle Digital Humanities*, *ivi*, p. 26.
[3] D. Fiormonte, *Per una critica del testo digitale. Letteratura, filologia e rete*, Roma, Carocci, 2018, p. 11.

cambiamento tecnologico rappresentano in quest'area. Capasso è stato Direttore dell'Istituto CNR di studi sul Mediterraneo, e adesso ricopre il ruolo di Direttore del Dipartimento Scienze Umane e Sociali, Patrimonio Culturale del CNR, su cui insiste l'area tematico-progettuale delle Digital Humanities, trasversale ai diversi Istituti che afferiscono al Dipartimento.
L'altra sessione è dedicata alla "Filologia digitale", e si apre con un intervento di Andrea Mazzucchi, dal titolo "Approcci integrati per la filologia digitale: l'esempio dantesco tra testo e paratesti", che intende mostrare attraverso alcuni recenti progetti quanto la transizione digitale in ambito filologico possa essere fruttuosa per restituire efficacemente la complessa testualità medievale e i meccanismi della sua tradizione, con particolare riferimento ai testi di Dante e all'esegesi antica della *Commedia*. Una prospettiva di grande rilievo, in quanto viene da un filologo della migliore e più autorevole tradizione italiana (che si identifica storicamente con la filologia dantesca), che però riveste anche il ruolo di Responsabile scientifico nazionale dello Spoke 3, "Digital Libraries, Archives and Philology" nell'ambito del Partenariato Esteso CHANGES, e coordina la realizzazione di un grande ecosistema integrato delle edizioni digitali dei testi capitali della letteratura italiana dei vari secoli. Nella duplice veste di filologo medievista e di *leader* di un progetto di così vasta portata egli offrirà il suo contributo e il suo sostegno scientifico al convegno.
Desideriamo ringraziare il Comitato di programma, i 10 Chair che hanno seguito la fase di review per gli assi tematici, e i 79 revisori che hanno generosamente contribuito (per un totale di 249 revisioni), tutti coloro che hanno risposto alla call for papers e i numerosi partecipanti iscritti al convegno. Un sentito ringraziamento va al direttivo dell'AIUCD e in particolare alla presidente Marina Buzzoni, alla vicepresidente Cristina Marras, al segretario Paolo Monella e alla tesoriera Francesca Frontini per il sostegno, il supporto e soprattutto per la disponibilità amicale che ci hanno dimostrato accompagnandoci in questi lunghi mesi di preparazione al convegno.
Last but not least, un ringraziamento speciale ai giovani della segreteria scientifica, per l'energia, l'entusiasmo e la determinazione con cui hanno lavorato, in armonia e amicizia, sul versante sia logistico che scientifico, per la realizzazione di questa importante iniziativa. Alle loro mete, umane prima che digitali, vorremmo dedicare questi atti che siamo onorati di poter curare e presentare alla comunità scientifica.

Antonio Di Silvestro
Daria Spampinato

RELAZIONI DEI KEYNOTE SPEAKERS

Informatica umanistica, Infocrazia, e Intelligenze Artificali[1]

Giuseppe Savoca
Università di Catania, Italia

Ringrazio l'Associazione per l'Informatica Umanistica e la Cultura Digitale e gli organizzatori di questo convegno per avermi voluto affidare l'apertura dei lavori.
Non sono un esperto di informatica e non ho particolari competenze o meriti sugli argomenti che costituiranno oggetto delle relazioni e degli interventi in programma. Sono stato forse quello che si poteva definire un utente informato, e cioè uno che ha intravisto i vantaggi di una introduzione dell'uso del computer nei propri studi di Italianistica. Posso dichiarare in tutta sincerità che la mia cultura e la mia formazione sono collocate nell'epoca del libro scritto, e anche la mia carriera di docente si è svolta nell'ambito, diciamo così, dell'Umanesimo cartaceo. Ora forse i docenti saranno sostituiti da lezioni programmate e fornite dall'intelligenza artificiale e gli studenti potranno fare a meno dei libri, e magari dei docenti in carne ed ossa.

SULL'INFORMATICA UMANISTICA A CATANIA

Alla metà degli anni Sessanta credo di avere capito che gli studi umanistici non avrebbero potuto avere un futuro se non si fossero aperti alle nuove possibilità offerte dagli strumenti di calcolo. Ovviamente eravamo in tanti a pensare che queste macchine non andavano riservate esclusivamente agli studi cosiddetti scientifici, ma costituivano una sfida e una opportunità per aprire in tutti i campi del sapere nuovi orizzonti di conoscenza. Era però difficile trovare colleghi italiani che passassero dalle buone intenzioni ai lavori concreti sul campo.
Ho allora tentato in autonomia pratiche di accostamento ai testi letterari assistite dal computer. E ho puntato da subito a costituire insiemi linguistici che fossero idonei a darci una figura dei testi poetici italiani. Pensavo che ogni testo potesse essere consegnato a una memoria non più cartacea, e di arrivare così a una condizione di lettura che potesse in tempi rapidi farci cogliere la trama lessicale e tematica su cui i testi erano nati e si erano formati. Padre Busa non lavorava in Italia al suo *Index Thomisticus*, e l'unico centro italiano che mostrava interesse per la conservazione elettronica dei testi era attivo in campo filosofico a Roma; promosso e guidato da Tullio Gregory, questo centro, incardinato poi nel CNR, si sarebbe chiamato negli anni Settanta "Lessico Intellettuale Europeo" (L.I.E.).
Per dirla in breve, mi è sembrata una strada percorribile quella di lavorare a un progetto di studio in cui il computer potesse essere un nostro alleato nella costruzione e nello studio del sistema lessicale di una singola opera, di alcune o di tutte le opere di un unico autore, e poi di più autori vicini per cronologia, genere letterario, secolo, scuola, ecc. La prospettiva di fondo era ed è sempre stata quella umanistica di partenza, e cioè cercare di cogliere la bellezza e il senso nascosti in ogni testo di autentica poesia.
In termini generali, l'esigenza che mi spingeva era di ordine critico e, diciamo, ermeneutico, e posso riassumerla con l'aforisma di Friedrich Schleiermacher che dice così: "Ogni comprensione del singolo elemento è condizionata dalla comprensione del tutto" (*Ermeneutica*, pag. 183).
Il tutto era/è ancora per così dire espandibile, e andava dal vocabolario-concordanza della singola opera (in sé compiuta) a tutte le opere di un autore, e poi alle opere di più autori affini per lingua di base, corrente letteraria o altro. Salto a piè pari queste problematiche, e aggiungo che ho definito questo metodo operativo adottando sin dall'inizio il concetto di **modularità.** Faccio un esempio. Ho elaborato e pubblicato (ma la pubblicazione è secondaria perché avevo comunque la banca dati sempre utilizzabile) due concordanze: una di tutte le poesie di Montale, e una di tutte le poesie di Ungaretti. Ognuna delle due può essere usata e interpretata autonomamente, ma esse rientrano anche in una banca dati più ampia, e sempre implementabile, in cui ci sono altre concordanze poetiche come un tutto potenziale in continua espansione.
Aggiungo, per essere più chiaro, e non certo per vanità autopromozionale, che, trattando con un programma specifico da me ideato le due concordanze, e cioè i due insiemi linguistici di Montale e di Ungaretti, io ne ho prodotti altri, tra i quali un insieme delle intersezioni lessicali tra i due poeti. Ho potuto quindi, tra l'altro, studiare contrastivamente il loro vocabolario specifico e quello che essi hanno in comune, mettendolo poi in rapporto con il lessico di Leopardi e di altri. Di questi studi ho poi dato qualche ragguaglio nel libro *Parole di Ungaretti e di Montale* (Roma, 1993). E come sintesi di sedici poeti italiani del Novecento, ricordo anche il mio *Vocabolario della poesia italiana del Novecento*, pubblicato da Zanichelli nel 1995.

[1] Il testo riproduce l'intervento tenuto in occasione del XIII Convegno AIUCD 2024 (Catania, 28-30 maggio).

In termini generali di metodo, osservo che i lavori appena citati sono tappe del progetto originario, e cioè moduli autosufficienti e insieme assemblabili. Senza mai avere studiato i linguaggi dell'informatica, ho poi scoperto di avere applicato nel mio lavoro sui testi un concetto tecnico e metodologico di modularità che era centrale nell'informatica di allora, e forse lo è anche in quella di oggi.
La "modularità" è propria dell'ingegneria del software, e si struttura intorno a un principio di progettazione e organizzazione in cui un sistema complesso viene suddiviso in parti più piccole e ben definite chiamate "moduli".
In realtà, il *Lemmatizzatore automatico* che ho realizzato ha un menù molto vasto in cui i singoli passaggi procedurali hanno una loro autonomia gestionale e una loro autosufficienza, per esempio relativa ai testi, alle liste di frequenza, all'individuazione automatica dei lemmi, alle categorie grammaticali, alle statistiche, ecc. Facendo lavorare in sequenza questi moduli, il risultato finale complessivo è stato quello di produrre il lessico integrale e specifico di ogni opera sotto la forma di un vocabolario esaustivo in cui i lemmi sono tutti classificati grammaticalmente e statisticamente. Ho discusso di questo metodo nel mio libro del 2000 intitolato *Lessicografia letteraria e metodo concordanziale*.
E mi scuserete se cito anche il *Vocabolario della poesia di Giacomo Leopardi*, uscito da Olschki nel 2010 come numero 26 della mia collana di "Strumenti di Lessicografia Letteraria Italiana". Ho dedicato questo libro al padre Roberto Busa, "pioniere dell'informatica linguistica, per i Suoi splendidi e saggi 97 anni (28 novembre 2010)". Gli avevo prima dedicato la *Concordanza di tutte le poesie di Guido Gozzano* nel 1984, e lui mi espresse più volte il desiderio di volermi dedicare un suo libro.
Tralascio gli aspetti privati della nostra lunga e totale amicizia, e dico solo che, avendo già costituito a Catania un Centro di Informatica Letteraria, feci chiamare ad insegnare da noi Linguistica computazionale proprio Roberto Busa. Le sue lezioni ci hanno illustrato splendidamente la bella storia della sua precoce intuizione di introdurre il calcolo elettronico nello studio dei testi e delle lingue.
Ancor prima di incontrarmi con lui, avevo organizzato una serie di incontri di studio con autorevoli colleghi italianisti, alcuni dei quali erano decisori dei finanziamenti nel settore umanistico presso il Consiglio Nazionale delle Ricerche. Nacque così un Gruppo nazionale di coordinamento del CNR che assunse il titolo di CLIPON, acronimo che sciolto si legge Concordanze della Lingua Italiana Poetica dell'Otto/Novecento. I volumi a stampa sarebbero stati poi pubblicati, sempre in fotocomposizione, dall'editore Olschki di Firenze nella collana di Strumenti di Lessicografia Letteraria Italiana, che è stata attiva dall'anno 1984 fino al 2011, e che costituisce un piccolo *corpus* di oltre 12000 pagine.
Il mio intento è stato anche quello di coinvolgere gli studenti, e poi i dottorandi del Dottorato in Lessicografia e semantica del linguaggio letterario europeo, nell'uso delle nuove tecnologie, e numerose sono state, sin dai primi anni Settanta, le tesi di laurea e di dottorato svolte con l'ausilio dei programmi da me elaborati a Catania con la collaborazione di informatici esterni all'università.
Indipendentemente da Busa e dall'allievo Zampolli, che avevano a loro supporto rispettivamente il gigante IBM e il CNUCE di Pisa, ho elaborato, perfezionandolo negli anni, un programma di spoglio integrale lessicografico dei testi, registrato poi alla SIAE come "Lemmatizzatore automatico per concordanze con banca dati del lessico della poesia italiana". Cito questo particolare perché, anche se gli umanisti non producono beni di consumo monetizzabili, il loro lavoro intellettuale si esprime in pubblicazioni in genere coperte da diritto d'autore. Queste opere appena stampate vengono acquisite, e sempre senza il loro consenso in grandi big data che i giganti del web e i grandi produttori di software (Google in testa) utilizzano oggi per i loro progetti di IA.
Fino a qualche anno fa si poteva, ad esempio, interrogare liberamente su Google Books qualunque mia concordanza. E qualcuno ricorderà che ci sono state proteste da parte dell'UE approdate infine a un accordo con Google. Oggi si vede online solo qualche frammento dei libri immagazzinati dal colosso del web. Ma non è verosimile immaginare che Google abbia distrutto l'immenso archivio accumulato per tutte le letterature di tutti i paesi.
Dopo Busa è stato chiamato a Catania come docente di Statistica linguistica Charles Muller (Strasburgo, 1909-2015), che era allora, e per me resta ancora, il massimo studioso della linguistica computazionale in chiave statistica. Busa e Muller, con tanti altri, nell'aprile del 1985 parteciparono a un convegno svoltosi in questa università su "Lessicografia, filologia e critica", che si occupava dei "Problemi della ricerca letteraria a confronto con le nuove metodologie e con l'informatica".
Il convegno fu aperto da una relazione di Busa su "Informatica e nuova filologia", a cui seguì una relazione di Muller su "Lexicologie, statistique lexicale et critique littéraire", mentre nel pomeriggio Antonio Zampolli fece una relazione sul "Dizionario di macchina dell'Italiano".
Muller tornò a parlare nel pomeriggio della "banca dati ortografici e grammaticali", e fu una bella novità il collegamento telematico attivato in tempo reale con questa banca dati che risiedeva a Strasburgo. Dico per inciso che allora non esisteva internet, e che le videoconferenze erano una rarità. Noi riuscimmo, via modem e con un terminale videotelematico chiamato in Francia Minitel, di cui non c'era l'equivalente in Italia, a interrogare la banca dati "Orthotel", per esempio, sulla coniugazione di un verbo francese.

Degli altri partecipanti attivi nel campo dell'informatica umanistica ricordo solo l'italo-americano Luciano Farina e Tito Orlandi. Orlandi era già stato, insieme a Giovanni Adamo, in visita al nostro Centro di Informatica Letteraria nel maggio del 1984, e di ciò che trovarono qui si può leggere il loro Rapporto nel volume *Dall'informatica umanistica alle culture digitali* (Roma, 2012, pp. 57-60).
Ma torno un poco al profilo di Roberto Busa, la cui opera è assolutamente centrale nel quadro dell'Informatica umanistica del Novecento.
Busa nel 1949 a New York presentò la sua idea di indicizzazione automatica del lessico di San Tommaso d'Aquino al presidente della IBM. Il colloquio sembrava non essere stato fruttuoso, ma padre Busa, uscito dall'ufficio del capo, prese un cartello con lo slogan IBM di allora, tornò indietro e glielo mise sotto gli occhi, dicendogli: «Ma non siete stati voi a scrivere che siete in grado di realizzare subito le cose difficili, mentre volete un po' più di tempo per le cose impossibili?». Questa battuta valse a Busa l'apertura di tutte le porte della IBM nel mondo. Va anche detto che il presidente non fu meno arguto di lui nella risposta, quando precisò i "limiti" della concessione nel patto che «IBM» non diventasse il marchio di «International Busa Machines».
Cominciò così un lavoro di collaborazione tecnologica che ha sperimentato sul campo e ha accompagnato tutti i progressi novecenteschi dell'archiviazione elettronica (dall'uso delle schede perforate ai nastri magnetici e alle memorie volatili). L'impresa è approdata alla stampa in fotocomposizione (completata nel 1980) dei 56 volumi dell'*Index Thomisticus*, un immenso vocabolario di 11 milioni di voci che, con le 118 opere di San Tommaso e 61 di scrittori medievali, contiene 20 volte il numero di righe dell'intera Enciclopedia Treccani. Il lavoro fu riversato in un Cd-rom nel 1992.
Parallelamente alla concreta pratica lessicografica e alla ricerca, padre Busa ha esercitato una vasta attività didattica e divulgativa, insegnando "Digital Humanities" all'Università Cattolica di Milano e alla Pontificia Università Gregoriana di Roma, e "Intelligenza Artificiale e Robotica" al Politecnico di Milano. Da un suo rapporto ai superiori del 2008, risulta la sua partecipazione a 245 congressi, con lezioni e corsi tenuti in tutti i continenti. Egli ha lavorato su 22 lingue e su 9 alfabeti, e la sua apertura ecumenica lo ha visto impegnato tanto sui testi biblici di Qumran quanto sul Corano.
Tra i suoi volumi ricordo solo i rivoluzionari *Fondamenti di informatica linguistica* del 1987, e i più "leggeri" *Quodlibet. Briciole del mio mulino* (1999) e *Dal computer agli angeli* (2000), costituito da "milleduecento momenti di pensiero". Quest'ultimo titolo è emblematico, direi, della visione religiosa e "trascendentale" che egli coglieva nella "creatività". Per lui "L'informatica è un frutto della creatività ed è la creatività che ne misura il progresso" (*Briciole* ..., p. 75). Nella dedica privata del Natale 2000 di questo libro leggo, sempre con timore e tremore, queste parole: "A Peppino Savoca che pure dalle parole va agli Angeli, espressioni del Mistero della Vita e del tempo". "Mistero", "Vita", "tempo" sono parole pesanti, ma Busa aveva, oltre la fede, l'ottimismo di una visione felice, e direi persino utopica, del mondo digitale. Il sottotitolo di questo libro *Dal computer agli angeli* si conclude sulla sua certezza di avere "inquadrato" "le reti elettroniche entro quelle degli spazi della vita".
Busa è stato ispiratore e consigliere di numerose imprese di raccolta, classificazione e studio di testi letterari, scientifici, filosofici e tecnici. Ha ideato e avviato progetti ancora da proseguire e studiare, come quello sulle "Lingue Disciplinate", che segna una svolta radicale nel difficile campo della traduzione automatica internazionale, che egli rappresentò come una sfida posta all'Unione Europea dalla globalizzazione delle comunicazioni in rete. Il suo lascito morale, il suo «Testamento» umano, come scrisse lui stesso, si colloca tra «profezia e utopia», ed è per tutti, ma soprattutto per i giovani. Queste le sue parole: «Chiudo la mia vita di lavoro contento e ottimista. Oltre che per motivi di ordine personale, il mio ottimismo proviene dal fatto che questa nuova metodologia apre ai giovani vasti campi di lavoro».

INFOCRAZIA

Passo al secondo punto di questa introduzione, relativo alla "infocrazia", e cioè al potere dell'informazione digitale nel mondo di oggi e di domani.
I sociologi e i filosofi, a partire dalla seconda metà del secolo scorso, si sono posti più volte il problema dei rapporti della rivoluzione telematica con la democrazia. Penso, ad esempio, al libro *La condizione postmoderna. Rapporto sul sapere* (1979) in cui Jean-François Lyotard immaginava, insieme ai pericoli di un uso perverso delle nuove tecnologie a vantaggio del capitalismo imprenditoriale, la possibilità di un allargamento dello spazio pubblico di controllo dell'agire politico. Egli vedeva chiaramente come il sapere, cioè la conoscenza in tutti i campi dell'umano, fosse la maggiore posta in gioco "della competizione mondiale per il potere" (p. 14). La sua era la visione positiva di una società in cui "il pubblico deve avere libero accesso alle memorie e alle banche dei dati" (p. 121). Alla fine il filosofo prefigurava "una politica in cui saranno ugualmente rispettati il desiderio di giustizia e quello di ignoto" (p. 122).
Sul fronte opposto, altri hanno motivatamente argomentato come il controllo delle banche dati contenenti informazioni sensibili su milioni di cittadini sarebbe diventato una fonte di enorme potere in mano a pochi.

Le due posizioni si possono rappresentare anche con due parole (due concetti) ricorrenti in letteratura, in informatica e in filosofia quali sono utopia e il suo contrario distopia. In mezzo ci stanno, ci starebbero, la scienza, la cultura, e anche, diciamo, l'informatica "libera", tra cui anche le Digital Humanities nella versione che chiamerei universitaria.
Torno a citare padre Busa per rilevare che egli si rese da subito conto che l'informatica umanistica non sarebbe potuta nascere, né, nata, avere uno sviluppo ad opera dei soli filosofi, linguisti o letterati. Da qui il suo gesto di rivolgersi, per fortuna con successo, al potere finanziario di una grande impresa multinazionale come la IBM.
E va anche ricordato che il pioniere dell'informatica linguistica ne ha intravisto da subito il risvolto di potere tecnicamente economico, politico e persino militare. Padre Busa, che era un uomo straordinariamente candido e come tale rispettato da tutti, ha operato attivamente, e in tutta tranquillità, sostenuto da suoi superiori, dal Vaticano, dalle imprese e da enti pubblici nazionali e sovranazionali.
Ricordo che egli si interessò tra i primi ai problemi della traduzione automatica. In questo ambito è entrato in contatto con diverse università statunitensi, partecipando ai tempi della guerra fredda ad un progetto di traduzione automatica dal russo all'inglese finanziato dal Pentagono, cioè dal ministero della difesa USA tra il 1952 e il 1966. L'ALPAC (Automatic Language Processing Advisory Committee = Comitato consultivo per l'elaborazione automatica della lingua) fu bloccato dallo stesso Pentagono, e Busa, che ha continuato a lavorare al settore fino al primo decennio del duemila, restò alquanto deluso. Cito alla lettera le sue parole: "Sfortunatamente, nel 1966, come risultato del Rapporto Alpac, il Pentagono tagliò tutti i fondi".
Su questa interruzione dei finanziamenti da parte del potere politico-militare osservo che i tempi lunghi della ricerca scientifica sono per lo più incompatibili con le esigenze di rapidità ed efficienza produttiva volute dal potere di turno. Ed è molto probabile che nel caso particolare il Pentagono abbia proseguito autonomamente i suoi programmi di traduzione automatica dal russo.
In buona sostanza, il potere dell'informazione è sempre in posizione di inferiorità rispetto ai poteri e agli interessi degli imprenditori, della finanza, della politica. Il termine "infocrazia" è stato coniato dal filosofo coreano-tedesco Byung-Chul Han; esso dà il titolo a un suo libro del 2021. Il sottotitolo è "Le nostre vite manipolate dalla rete", ed è alquanto espressivo delle problematiche oggi dibattute tra big data, intelligenze artificiali e futuro delle società e dello stesso uomo.
La parola "infocrazia", che significa potere legato all'informazione digitale, rientra in un campo lessicale molto vasto in cui c'è, ad alta frequenza, anche "democrazia".
Parole dello stesso ordine di significato, e forse anche più frequenti di "infocrazia", sono, tra le altre, "algocrazia" (cioè potere degli algoritmi) e "datacrazia" (cioè potere dei big data).
Se dovessi inquadrare l'infocrazia di Byung-Chul nell'opposizione prima accennata tra utopia e distopia, direi che essa inclina decisamente dalla parte della distopia. Essa infatti descrive come già in atto una deriva della democrazia digitale (vista in positivo da altri) verso l'esito di un mondo intrappolato in una caverna digitale in cui non risplende più la luce della verità. Per il filosofo il nuovo nichilismo del XXI secolo è segnato dalla perdita del contatto con la realtà e con le verità fattuali.
Secondo lui dobbiamo prendere atto e piena consapevolezza che siamo ormai nell'era delle *fake news*, le quali "non sono menzogne: esse attaccano la fatticità stessa", e cioè attaccano quello che Habermas chiamava il "mondo della vita".
La conclusione amara, su cui credo che ognuno debba attentamente meditare, è che "La democrazia non è compatibile con il nuovo nichilismo. Essa presuppone un parlar vero. Solo l'infocrazia può fare a meno della verità" (p. 73).

AUTOMI E INTELLIGENZE ARTIFICIALI

Chi si occupa di informatica oggi non può prescindere dall'affrontare il tema degli automi su cui abbiamo molte notizie già nelle letterature dell'antichità: ad esempio a partire dalla cameriera automatica immaginata dall'ingegnere del terzo secolo avanti Cristo Filone di Bisanzio. E come antenata dei robot si ricorda anche la Macchina teatrale realizzata da Erone di Alessandria forse nel primo secolo.
Numerosi sono i robot dell'epoca moderna. Si citano, ad esempio il Cavaliere meccanico di Leonardo da Vinci (attorno al 1495), il Monaco del XVI secolo che camminava e pregava, un androide che suonava il flauto costruito nel Settecento, un'anatra meccanica, una tigre e altri animali. Saltando all'oggi e al domani, accenno al robot TongTong, che è la "bambina" robot che i giornali di questi mesi ci dicono come creata dall'intelligenza artificiale in una Cina con famiglie senza figli. Essa avrebbe un "sistema mentale e di valori di una bambina di 3 o 4 anni", ma con la possibilità di imparare dall'esperienza, e questo al fine di assistere sempre meglio gli anziani.
Credo tuttavia che sia già adulto l'automa chiamato Sophia che, come ci informano le cronache, terrà il discorso finale ai laureati dell'università di Buffalo negli USA.

Com'è noto, un automa può essere tanto una macchina fisica che imita il corpo e i comportamenti umani, quanto un programma immateriale e invisibile che elabora informazioni, opera, risponde a domande, impartisce ordini, e in fondo dispone della nostra vita.
In questa seconda attività rientra il concetto ampio, e ancora in corso di definizione, di intelligenza artificiale. In un mondo in cui le grandi imprese del web sono in conflitto non dichiarato, ma reale, tra di loro si potrebbe arrivare a un trionfo dispotico di pochissime IA, e a un mondo abitato da robot umanoidi, da uomini deumanizzati e da animali resi quasi umani.
Nei discorsi sull'inquinamento globale capita di imbattersi nell'ipotesi che, chiedendo all'IA la soluzione del problema, questa lo indichi seccamente nella eliminazione o nell'estinzione dell'umanità. È di questi giorni la notizia di uno studio che, in relazione al progressivo riscaldamento del pianeta, prevede che tutti i mammiferi scompariremo entro 250 milioni di anni. Ma forse questo potrebbe accadere molto prima. E non bisogna sottovalutare il fatto che oggi, e sempre di più, il maggior consumo di energia è imputabile ai giganti del Web e ai supercomputer quantistici.
Abbiamo appena letto nei quotidiani che in Italia il 19 maggio si è consumata la somma delle risorse naturali (foreste, acque, ecc.) che noi italiani avremmo disponibili per tutto l'anno 2024. In termini numerici ciò significa che in meno di 5 mesi si è esaurito quello che la popolazione italiana dovrebbe ricevere dalla natura per l'intero anno. Per i restanti 7 mesi noi vivremo rubando alla natura risorse non rimpiazzabili. La conclusione di questo saccheggio è che noi siamo in conflitto con la natura. E questo accade su scala mondiale.
È ovvio che questi dati e queste previsioni sono oggi possibili perché un algoritmo ha trattato una serie vastissima di informazioni organizzate in una banca dati specializzata sull'ambiente, sui consumi, ecc.
L'IA può oggi fare questo e tanto altro, in tutti i settori e gli ambiti della vita, della società, del mondo.
Cito il più noto scrittore di fantascienza tecnologica che è Isaac Asimov, e in particolare il romanzo *Io, Robot* del 1950, in cui la prima legge della robotica viene enunziata in questi termini: "Un robot non può ferire un essere umano o, per inerzia, permettere che un essere umano venga danneggiato".
Questa legge è di grande attualità perché pone lucidamente il problema della sicurezza degli uomini rispetto ai possibili impieghi distorti di macchine robotiche e, aggiungiamo, di qualunque programma di IA.
L'ultimo capitolo di *Io, Robot* si intitola *Il conflitto evitabile*, e già nel titolo ci rivela la direzione utopica del pensiero di Asimov, che, prima della rivoluzione informatica dovuta appunto all'IA, prefigurava un mondo governato da un'etica dei rapporti sociali e dell'uso della cibernetica in cui l'uomo fosse sempre fine e non utente passivo, lavoratore inconsapevole, e non retribuito, a servizio dei grandi monopoli digitali e di poteri politici che sfuggono a ogni controllo dal basso.
E dunque il discorso sulla IA non può che tornare al tema del potere, della forza, del conflitto e infine del nichilismo insito nella infocrazia.
Noi viviamo ancora nell'era atomica, e siamo sempre sotto la minaccia di una fine dell'umanità dovuta a un conflitto nucleare che, per scelta o per errore di qualcuno, potrebbe incombere sul nostro presente. C'è da chiedersi se sia possibile che l'IA, la quale, in linea costitutiva, riguarda tutto il conoscibile, non riguardi anche questa realtà di catastrofe ultima.
In verità, le armi e la guerra sono molto probabilmente i campi privilegiati in cui oggi si concentra il massimo impegno dei poteri statuali e degli eserciti che lavorano al potenziamento dell'IA.
La stampa ci dice che sono in corso contatti diplomatici fra le superpotenze al fine di ridurre i rischi legati alle armi nucleari guidate dall'IA e di eliminare la possibilità che un singolo potente possa premere il bottone dell'apocalisse.
Ma l'IA è già da tempo al servizio della guerra e della morte. Richiamo ancora il fatto, già ricordato a proposito di Busa, della partecipazione del Pentagono al progetto ALPAC sulla traduzione automatica.
Il convegno di oggi è centrato sul Mediterraneo, sui suoi contesti e sulle sue reti. Allargo un attimo il concetto di rete per dire che noi possiamo comunicare digitalmente e usare banche dati perché c'è una enorme rete satellitare che ce lo permette, ma c'è anche una rete di cavi sottomarini che attraversa il nostro mare e in cui passa una buona percentuale del traffico internet mondiale.
Entrambe queste reti (come quella del Mar Rosso minacciata dagli Houti) sono possibili obiettivi militari, con le conseguenze che vi lascio immaginare. Aggiungo anche che, a reti fisiche inviolate, il sistema internet e tutto il mondo del digitale possono essere oggetto di attacchi capaci di rendere impossibili i normali funzionamenti di istituzioni, apparati e servizi essenziali.
Ed è anche molto verosimile che gli hackers siano funzionali a diversi poteri e che l'IA sia per loro una grande opportunità.
Le cronache di questi mesi relative ai conflitti in corso in Europa e nel Mediterraneo ci informano, per quello che è possibile sapere e dedurre, sugli usi anche terroristici e bellici delle risorse informatiche. In generale va osservato che, oggi, l'IA è, e diventerà sempre di più, il campo di conflitto tra le *intelligences* che guidano e sostengono le forze combattenti sul terreno.
E questo almeno fino a quando i soldati non saranno sostituiti dai robot.

Su Sky Tg 24 dell'11/12/2023 è stata data questa notizia: "L'esercito ebraico utilizza l'IA per individuare e bombardare gli obiettivi dei terroristi all'interno della Striscia. Ma dato il numero di civili palestinesi rimasti uccisi, alcuni analisti internazionali nutrono dubbi sulla sua accuratezza."
Sono state le stesse forze israeliane a rivelare che una delle loro IA, chiamata The Gospel (Il Vangelo), è un sistema che "consente l'utilizzo di strumenti automatici per produrre target a ritmo rapido". Questo significa che l'esercito israeliano ha raccolto in un immenso archivio i dati personali, biometrici, i movimenti ed altro su persone ritenute pericolose, e significa che questi dati vengono affidati all'IA, la quale suggerisce e decide gli obiettivi umani da colpire.
In termini numerici sono calcolati anche i danni collaterali: per ogni obiettivo da colpire è stimato che altri 15/20 individui possano venire eliminati. Quando l'obiettivo da eliminare è molto importante, il numero di vittime accidentali può salire fino a 100. Il *Corriere della sera* (riprendendo il *Guardian*) ci informa che questa IA chiamata Gospel è stata definita "una fabbrica di omicidi di massa".
Sempre sul *Corriere della sera* (6 aprile 2024) leggiamo anche che i soldati dedicano «meno di 20 secondi a valutare l'eventuale obiettivo» e «si riducono a mettere il timbro di approvazione alle scelte» dell'Intelligenza Artificiale.
La verità pura e semplice è che le imprese dell'IA sono in forte competizione tra di loro, come le grandi potenze sono in lotta sotterranea per il predominio in tutti i settori della Infosfera.
Per dirla in breve, qualunque hacker futuro potrà distruggere un'intera biblioteca digitale, ma il libro continuerà a esistere. E forse il critico più bravo sarà quello che saprà cogliere nei big data e negli abissi di internet le cose essenziali e saprà consegnarle alle pagine di un libro.
È sperabile che siano soprattutto i giovani a trovare nel tempo del conflitto le forze per resistere alla distopia e per creare uno spazio e un tempo di salvezza.
E forse sarà la bellezza a salvare il mondo.

The medieval Mediterranean in…data? Interpretation, conjecture and digital methods

Tara L. Andrews
Institut für Geschichte, Universität Wien, Austria

One of the more humbling experiences of my career combining computation with historical studies was when I was invited to be on a panel for users of cloud computing services in Switzerland and found myself sitting next to a scholar from CERN, the huge particle physics lab. This was during the time when the keyword "big data" was bandied around a lot, including in the humanities, and specifically in the Digital Humanities where it seemed that a lot of the computational analysis of the "humanities computing" era had fallen by the wayside, in favour of the re-mediation and generative work that characterised the second wave (Presner, Schnapp, and Lunenfeld 2009; Berry 2011; Hayles 2012). On this panel, it couldn't have been clearer – and I think we all have long since accepted – that what we have in the humanities is nowhere near the volume of what other fields consider "big data".

If I may make a visual metaphor: we can imagine the solar system as representing all the data in the world, with particle physics represented by the Sun. The humanities are the four inner planets; historical studies might be the Earth, and medieval history must content itself with, approximately, the moon. Moreover, when looking at medieval Mediterranean societies, we almost always need to take a comparative approach. To continue to stretch the metaphor, this is approximately like trying to put together information about the chemical composition of Jupiter's gas cloud together with the atmospheric composition of the Moon together with observations from the middle of a hurricane on Earth. Meanwhile, we observe that Saturn has rings and wonders what this might tell us about its fellow planets.

This problem of lack of information, especially in the context of the need to do comparative analysis, is not limited to digital data per se; it is one of the challenges of doing medieval Mediterranean history at all. Indeed, we constantly hear (and make) these laments. As an experiment, I arbitrarily picked a volume off my shelf, which happened to be the edited collection *Social Change in Town and Country in Eleventh-Century Byzantium* (Howard-Johnston 2020). Every single article had some allusion, in one form or another, to the difficulties posed by lack of evidence concerning the topic of their paper. Perhaps the most striking passage is actually a form of counterexample, in an article by Tim Greenwood discussing what we can discern about the history of the province of Taron after its incorporation into the Byzantine Empire in the late tenth century. His reconstruction is based largely on written histories that were produced at that time in Taron, of which the portions dealing with contemporary history have been lost, so that the versions that come down to us only discuss events in the fifth to seventh centuries. Greenwood makes some very convincing suggestions by means of the detective work and suggestive comparisons, themselves based on fragments and scraps of evidence, that we must so often undertake. And yet, he begins his article thus:

"The social history of tenth- and eleventh-century Armenia has attracted little in the way of sustained research or scholarly analysis. Quite why this should be so is impossible to answer with any degree of confidence, for as shall be demonstrated below, it is not for want of contemporary sources." (Greenwood 2020, 196)

This passage illustrates that, by now, we are so accustomed to doing our work with such a dire state of surviving evidence that it is easy to stop noticing that fact. I encounter new and impressive works of scholarship all the time whose authors have taken closer looks at the evidence we do have, noticed assumptions that went unquestioned or unremarked for too long, and gives us really compelling new interpretations of the evidence that we do have – all this without the luxury of the "big data" of our colleagues at CERN, nor even the comparative abundance that the modern and contemporary historians can work with.

How, then, are we meant to make any progress in this field in the context of the Digital Humanities? The first step has always been collection of the evidence we have, in whatever form we have it, in a systematic way. Such collections have a long analogue history in our field. We can think of text editions with their apparatus criticus; we can think of concordances; we can think of prosopographies and gazetteers and manuscript catalogues. Indeed, scholars long ago started bringing resources such as these into the digital realm. Digital editions of ancient and medieval texts abound by now; digital prosopographies include the two major Byzantine projects *Prosopographie der mittelbyzantinischen Zeit* (PmbZ) (Lilie et al. 2013) and *Prosopography of the Byzantine World* (PBW) (Jeffreys 2017), the plethora of projects offered by Kings College London on a similar technical basis as PBW (Nelson and Tinti 2006; Mouritsen et al. 2015), and many smaller projects either finished or underway. In addition to this are resources such as the Pleiades gazetteer, the Pinakes manuscript

catalogue, and the various sigillographic and numismatic collections maintained by Dumbarton Oaks and others. Each of these is based on a set of data models, data standards, and frameworks in order to function. Some of them (such as those that rely on text encoding or geographic description) feel reasonably mature, and others (such as prosopography or, perhaps, manuscript description) still tend to give rise to custom data models.

Let us look more closely at prosopographies as an example of how we have been moving this data online. We can compare DPRR and PBW, two projects working off more or less the same data model and produced in the same department, to see how they overlap and how they differ. They are both built around the factoid model; that is, the data is essentially a collection of source statements indexed by the person about whom the statement was made. They have both defined a set of categories of information that can be collected about a person; in each case these categories are adapted to the information that is available from the sources they use. In the case of DPRR, these are limited to life dates, legal or social status, offices held, and relationships; both primary sources and secondary sources are used, and dates are attached to the items in the 'life dates' category and the 'career' category. In the case of PBW, there are around 15 categories of information, such as Dignity/Office, Death (but not Birth), Location, Authorship, Education, Ethnicity, Languages Spoken, Occupation, Description, and 'Narrative', which is a catch-all for statements made about a person that do not fall into any of the more formulaic categories. Factoids are taken exclusively from primary sources, and the only information with dates attached are the factoids in the 'Narrative' category.

The creators of both prosopographies were conscious of the need to disambiguate the information they included, and to make associations with external references where they exist. DPRR includes, where applicable, the numbering used in the Pauly-Wissowa encyclopedia. PBW had no such existing resource to use for external reference to its people, and its own numberings have been adopted by a few other projects; at the same time, they have included links to the Pleiades and Geonames gazetteers for the places they mention.

Both prosopographies have alternative indexes alongside the person search; in the case of DPRR these are the 'fasti search' and the 'senate search', allowing the retrieval of all known holders of a particular office or all known senators in a particular year. In the case of PBW the other major index is the 'seals module', which is a collection of information drawn from as many catalogues as possible about surviving lead seal specimens from the period of coverage.

It is also worth noting that both projects make their source code and data openly available via Github and/or a SPARQL endpoint, so that scholars can reuse it. In short, both projects have paid careful attention to findability, accessibility, and reusability of their data and have amply documented their data model and collection methods in associated publications, and in this sense are very good examples of what digital history projects should look like.

Herein, however, lies a conundrum: the online prosopography I hear spoken of approvingly most often is neither of this, but is rather PmbZ. For what it is, PmbZ is very well executed; the information in the print version has been indexed, cross references have been linked, and a faceted search has been made available. Some of the information about a given person has been drawn out into a data table at the top, but the vast majority of the information is in the 'lemma', which is essentially a small scholarly article with references, explaining the life, career, and significance of the person. This is, then, essentially a traditional form of scholarship cast into the digital medium, and it turns out that this is what most scholars are comfortable with. They compare PBW's factoids unfavourably with PmbZ's articles, even though the information in the latter is presented in a much less analysable and machine-readable form and with no API to speak of. The preference of so many scholars for PmbZ has been a constant source of frustration to the creators of PBW, and we might wonder why and how this has come to be so.

From my conversations with various scholars on the topic, the crux of the issue seems to be that scholarly prose is simply better at expressing where nuance and ambiguity lie, when we try to understand what we think happened in the past. The factoid model is perhaps too impartial; users would prefer to have the value judgment from a scholar they can trust about which information they should accept as true, as opposed to an interpretation-free list of what was written in the available primary sources. The purpose of PmbZ is to make this human-readable and interpretation-laden information accessible on the Internet; they aren't really thinking about computational analysis at all. One purpose of PBW, on the other hand, was always to express the information in a way that is structured enough to lend itself to some forms of computational analysis, which tends to require that ambiguity and openness to interpretation be kept to a minimum. How do we reconcile these competing objectives? If we cannot, then many medieval historians will see little value in digital methods and little incentive to use them.

This is a problem that many people have wrestled with over the years; most of them have adopted an approach that can be summarised as "make a version of the data that is as complete and accurate as possible according to my (or our) current understanding." A fine example can be found within the PBW project, specifically the subsection that deals with Byzantine lead seals. Its creator, Michael Jeffreys, observed that the readings and interpretations of these seals are known through a

combination of catalogues of wildly varying age and accessibility, and articles published about individual seals that refer back to these catalogues, but about whose existence non-sigillographers might have no idea. Jeffreys aimed to collect and present, as data, all the information that he and his colleagues on the project could locate and reconcile about the seals that had been either published in catalogues or examined in separate articles, exactly so that a historian who wished to use the data would have it on hand in a single place (Jeffreys 2009).

A similar approach has been taken for a different prosopography project in the making, the *People of High Medieval Georgia*, carried out by my former Ph.D. student James Baillie at the University of Vienna. He also faced the problem of reconciliation of conflicting information about the period under study and the people who lived there, and wished to express as much of this information as possible in a machine-readable way that could be used for further analysis. His solution was to settle on a single consistent reading of the data and discuss the justifications for his decisions, when the sources conflicted, in explanatory notes attached to the records (Baillie 2021). In this sense Baillie's work went beyond the aim of the PBW seals module in that it provides a guide for the user of where conflict or ambiguity exists, even if the nature of that conflict and ambiguity remains invisible to the computer algorithm that might make use of the data.

I would argue, however, that much of the problem lies in the data models we use. Our vocabularies for historical information gathering – be it about people, places, text or anything else – tend to be straightforward. They model the way we are used to thinking about the information we want to know, which is as a set of facts that we may or may not have. The trouble we face is that, in the digital domain, software doesn't cope well with information that is straightforward and ambiguous at the same time.

The easiest space to recognise the problem is the use of Linked Open Data (LOD) in the humanities. On the one hand, LOD is a popular model for collections of information in all sorts of domains. Its ability to be backed by formal ontologies allows for a very explicit modelling of any problem domain where it is applied, and LOD-based data collections are becoming increasingly used within the humanities (Oldman, Doerr, and Gradmann 2015). On the other hand, LOD was designed without controversy in mind, for the most part, and so these bits of information tend to get dissociated very quickly from the context in which they should be regarded. This is a problem that is recognised throughout the cultural heritage sector, especially in museum settings, wherever LOD collections have been developed (Kahn and Simon 2020), and it has clear applications to collections of information about medieval history.

This brings me to the STAR (Structured Assertion Record) model, which was developed to address exactly these issues of how we represent ambiguity and conflict in our data, and whose implementation is at the core of the RELEVEN project that I am leading until 2026. If we consider a typical LOD statement as in Figure 1, we know that a typical historian will want to know two things about this statement: who is claiming this, and on what basis? Our approach is a reification-based one: the predicate becomes its own entity so that the subject, predicate and object can be attached to a new entity called an assertion. When we have such an assertion, we then attach authority (who is claiming this) and provenance (on what basis) to it to complete the legs of our STAR.

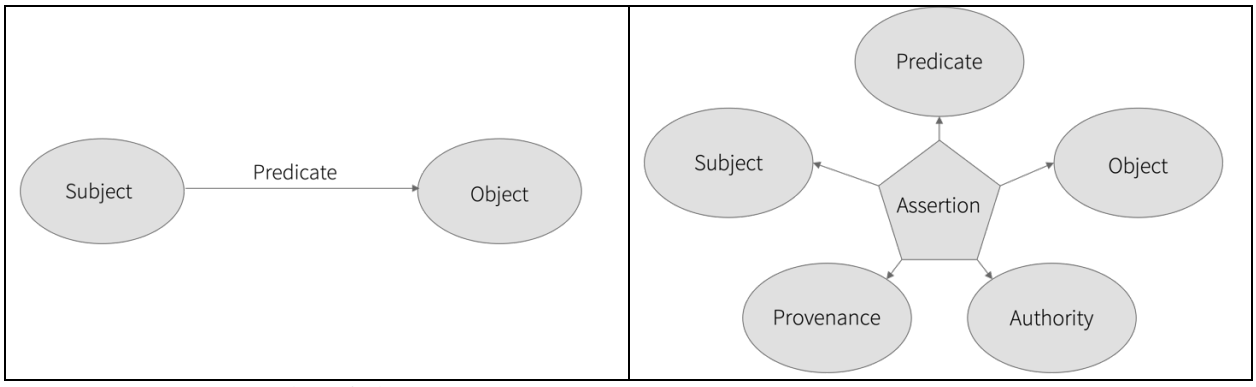


Figure 1. A generic LOD triple, reified into a STAR assertion

Our first task was to find how this approach can fit into existing models. As it happens, an argumentation model based on the CIDOC-CRM, known as CRMinf, was published in preliminary form already in 2015 and has had further development since (Doerr et al. 2023). Although CRMinf is quite complex and has yet to see much use, a relevant revision to the core CIDOC-CRM was published in 2019 which provided a class, `E13 Attribute Assignment`, through which we can implement the assertion concept, as seen in Figure 2. The use of this structure has led us to seek out CIDOC-CRM compatible vocabularies for the historical content that we are using; these include the base CRM itself, LRMoo (the successor to FRBRoo) to model the source material (Bekiari et al. 2022), and an adaptation of the SDHSS ontologies

(Beretta 2024) for their vocabulary concerning activities and relations between people. We are also at work right now on vocabularies for the expression of things such as journeys, epistolary exchanges, and political power structures over spatial extents.

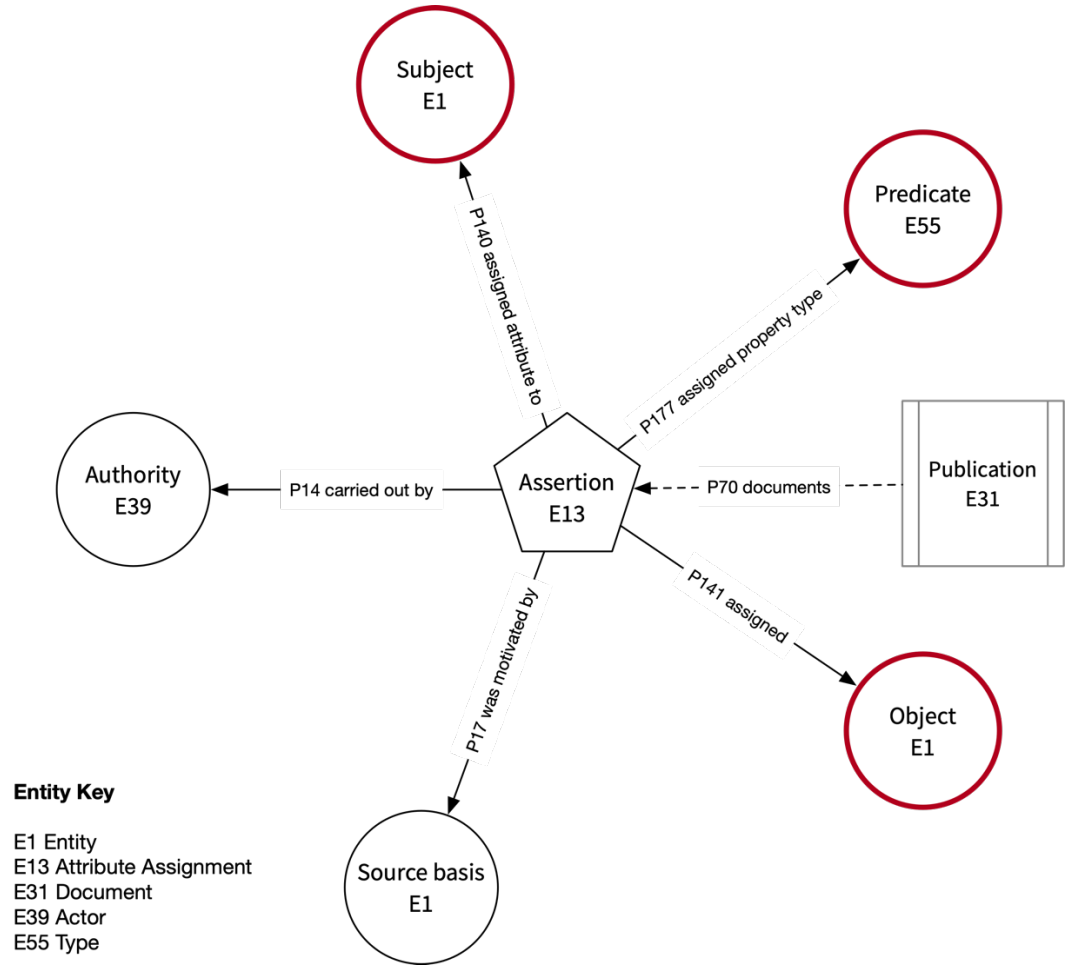


Figure 2. The STAR assertion model expressed in CIDOC-CRM

Figures A-Z provide an example of the model in practice. The statement together with the authority and the provenance (if it exists!) constitutes an assertion whose basis is a CIDOC-CRM statement, with the indication of who said it and whence it comes. In this case we have the 12th century historian Matthew of Edessa telling us (through his Chronicle – this means that we have documentation for the claim, but no specific idea of his sources for the statement) about the death of a Frankish mercenary called Frangopoulos, at the hands of the Byzantine emperor who was his employer (Matthew of Edessa 2017, 54). We represent here the date and the manner of his death.

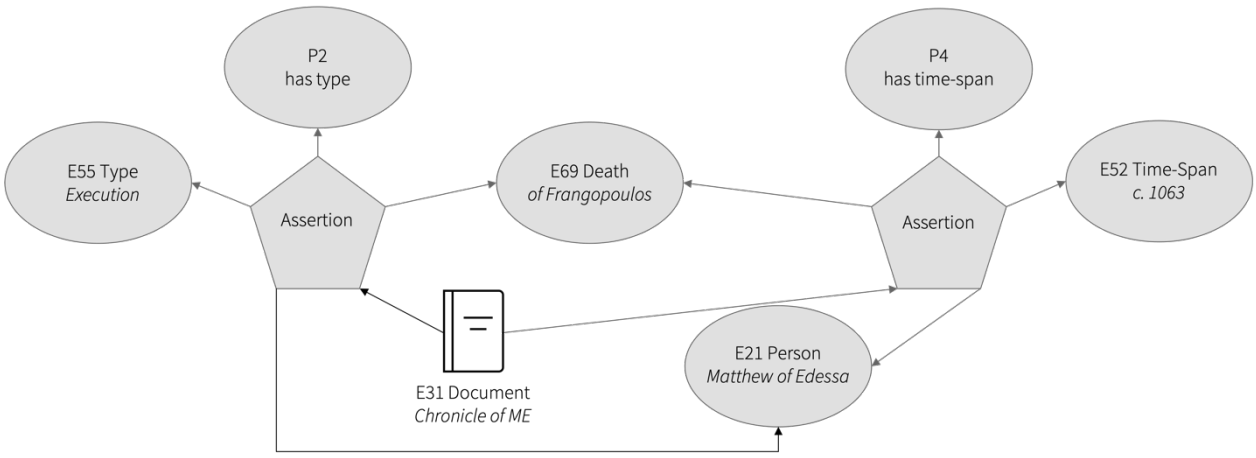


Figure 3. The death of Frangopoulos according to Matthew of Edessa

Now we come to a conflicting assertion based on a different source. The case has been made, based on a surviving Byzantine lead seal, that Frangopoulos was still alive and assuming (or at least trying to assume) command of the Byzantine armies in the east after their loss at the battle of Manzikert in 1071 (Seibt 2010). Just as in the factoid model used by most of the prosopographies mentioned above, it is up to the user to decide what to trust; so far there is not much difference. The question then becomes, how can we help users of our data collection identify where such conflicts arise, so that they are not left to work through the implications of each separate statement in isolation?

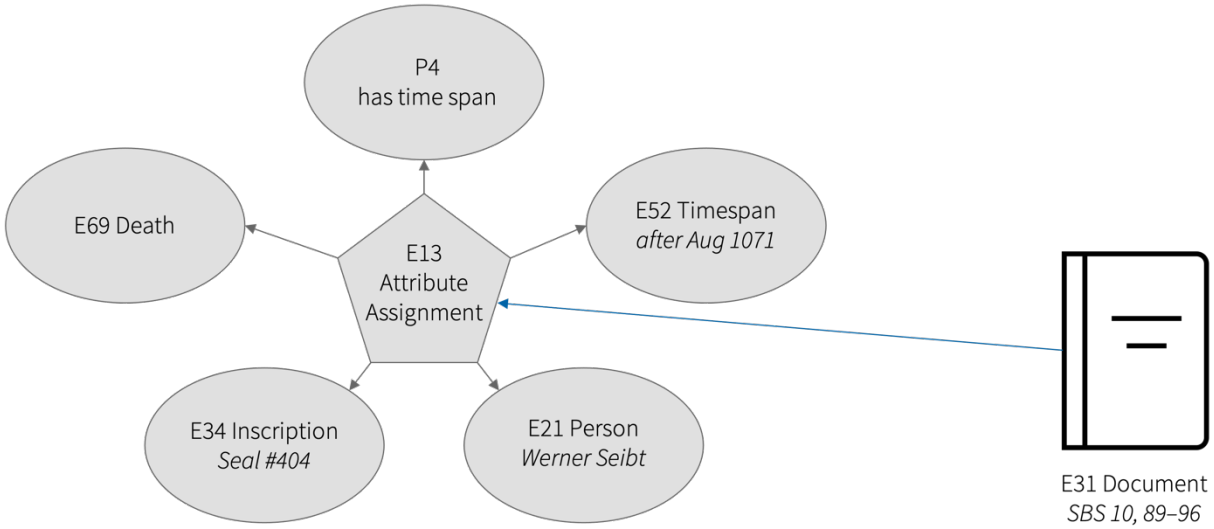


Figure 4. The death of Frangopoulos according to Werner Seibt (2010)

In a recent article in the realm of textual scholarship, Georg Vogeler made the passing complaint that, as much as we in the humanities are jumping aboard the ontology bandwagon, hardly any of us (except for a few research groups in Italy I can think of!) are making any use of inferencing (Vogeler 2021, 80). It seems that, in a way not dissimilar to the XML encoding of texts, for many of our colleagues the fun is in making the model, and others are left to actually use it. Detection of conflicts among assertions, however, is a place where it would make a lot of sense to use ontological inferencing. The aim would be to answer the question, which of the asserted triples cannot coexist in the same graph without a logic violation? And if we nevertheless believe all the triples are accurate, which of our logical assumptions are faulty? This is an approach that not only would make an interesting new use of ontological reasoning in a humanities context, but would also be a fine example of how modelling in the humanities can and should be used, as described by some of the pioneers of the field over 20 years ago (McCarty 2004; Unsworth 2002, referencing Orlandi 1999). I have an MA student working now on a prototype of exactly this problem, as it pertains to assertions of kinship between people in our various sources, and I am very much looking forward to the result this autumn.

We are also pushing the modelling a little bit further than single assertions. Let us come back to our eleventh-century mercenary Frangopoulos and the dissenting opinions about when he died. Our next modelling project concerns how we aggregate sets of assertions. There are two ways an aggregation might make sense; first would be the association of a single source statement, e.g. the (paraphrased) claim by Matthew of Edessa that "in the Armenian year 512 the emperor recalled Frangopoulos to Constantinople and had him executed by drowning", with the several CIDOC-CRM assertions that are necessary to model it.

The second would be to represent collections of assertions as viewpoints – that is, coherent pieces of world-view expressed by particular people at a particular time. In this case Matthew's world-view, as expressed in his *Chronicle*, includes the idea that the emperor had Frangopoulos executed around 1063 as punishment for his responsibility for the death of another military commander; Seibt's world-view as expressed in his article, on the other hand, has Frangopoulos still alive and attempting to take control of a chaotic situation in late 1071. We can model this, roughly, as in figure B. Such a model gains us two things: first, we can record (or perhaps our inferencer can deduce) that Seibt's own assertion necessarily implies a rejection of Matthew's dating; second, we can leave open the question of whether Seibt, or anyone else, would eventually accept Matthew's account as evidence that Frangopoulos was indeed recalled and executed by the emperor in place after 1071. In this way we are able to record certain forms of negative information, as well as explicit ambiguity, in our data model.

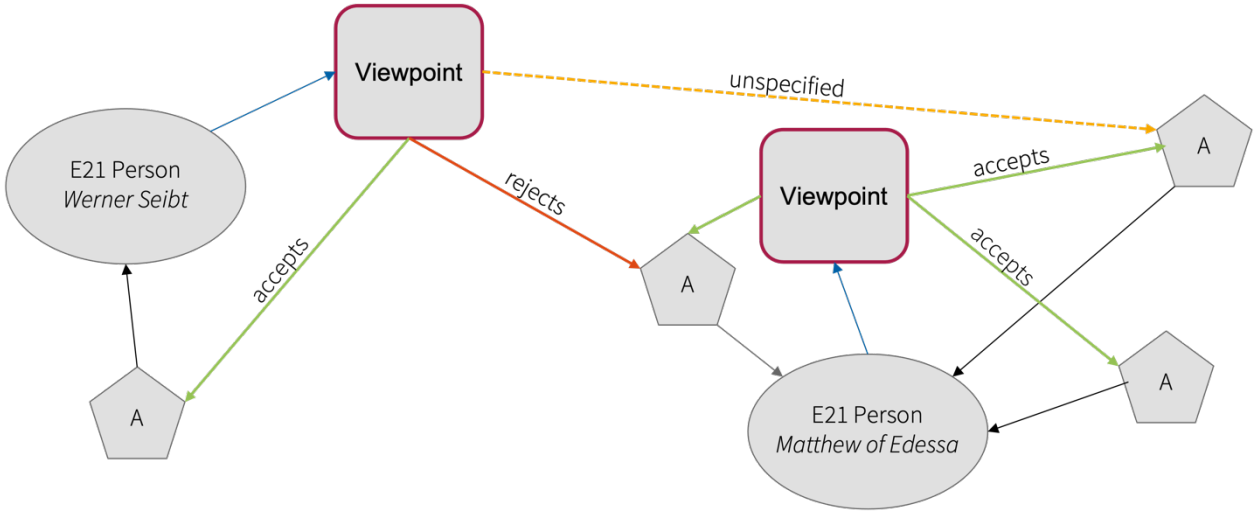


Figure 5. The viewpoints of Matthew of Edessa and Werner Seibt concerning the death of Frangopoulos

Such an extension of the STAR model brings us back to the question of existing models. Here the CRMinf extension re-enters the picture; we are currently in the process of determining whether this vocabulary meets our needs for viewpoint or statement expression. Much of the question about its use comes down to a sort of feasibility. CRMinf has a few extra layers of representation between the assertion and the viewpoint, which makes the model more complex than our initial sketch and may (or may not!) render it infeasible. A question also remains about how to model rejections of information. LOD works on the basis of positive subject/predicate/object statements; CRMinf makes it possible to claim that a particular statement is positively rejected by someone, but this requires us to instantiate the statement in order to represent its refutation and leaves open the question of how, and whether, a statement nobody holds to be true actually belongs in the database.

We must also be careful to avoid imagining that a model like this is a grand solution to all our data representation problems. It has already been show that even with CRMinf, it is difficult or impossible to capture the ambiguity, the rhetorical hedging of bets, that accompanies quite a lot of scholarly argumentation (Hacıgüzeller, Taylor, and Perry 2021). As I see it, the best we can do for the foreseeable future is use these tools to model specific and well-defined problems such as chronology, or kinships. And yet, that is still a great advance on trying to do this by hand!

To conclude, I'd like to step back from what we are doing in the RELEVEN project and focus on what is perhaps its main purpose. If we aim to work with the history of the medieval Mediterranean through the lens of data and computational analysis, I'd love to see us all stop focusing on "the fact" in our data models and start focusing on "the argument" as our primary unit. We need to stop thinking in terms of statements such as "when did Michael Psellos die", and be much stricter and more rigorous in thinking in terms of "what are the possibilities for when Michael Psellos died, and what evidence supports those possibilities?" Doing this will enable us to highlight conflicts and uncertainties in our sources instead of eliding them for the sake of the data model. It will allow our users to understand why scholars hold the opinions they do, and let the users carry forward their own work accordingly. And it may be a way to release more of the knowledge captured in scholarly prose into a domain where all these amazing methods of data science can do something useful for us.

REFERENCES

Baillie, James. 2021. 'Alternative Database Structures for Prosopographical Research'. *International Journal of Humanities and Arts Computing* 15 (1–2): 117–32. https://doi.org/10.3366/ijhac.2021.0265.

Bekiari, Chryssoula, Martin Doerr, Patrick Le Bœuf, and Pat Riva. 2022. 'LRMoo (Formerly FRBRoo) Object-Oriented Definition and Mapping from IFLA LRM'. https://cidoc-crm.org/frbroo/ModelVersion/version-0.9-0.

Beretta, Francesco. 2024. 'Semantic Data for Humanities and Social Sciences (SDHSS): An Ecosystem of CIDOC CRM Extensions for Research Data Production and Reuse'. In *Professorale Karrieremuster Reloaded. Entwicklung Einer Wissenschaftlichen Methode Zur Forschung Auf Online Verf{ü}gbaren Und Verteilten Forschungsdatenbanken Der Universit{ä}tsgeschichte*, edited by Thomas Riechert, Hartmut Beyer, Jennifer Blanke, and Edgard Marx, 73–102. Forthcoming. https://doi.org/10.33968/9783966270502-05.

Berry, David M. 2011. ‘The Computational Turn: Thinking About the Digital Humanities’. *Culture Machine* 12 (0). http://www.culturemachine.net/index.php/cm/article/view/440.

Doerr, Martin, Christian-Emil Ore, Pavlos Fafalios, Athina Kritsotaki, and Stephen Stead. 2023. ‘Definition of the CRMinf: An Extension of CIDOC-CRM to Support Argumentation’. https://cidoc-crm.org/crminf/sites/default/files/CRMinf%20v1.0%28site%29.pdf.

Greenwood, Tim. 2020. ‘Social Change in Eleventh-Century Armenia: The Evidence from Tarōn’. In *Social Change in Town and Country in Eleventh-Century Byzantium*, edited by James Howard-Johnston, 0. Oxford University Press. https://doi.org/10.1093/oso/9780198841616.003.0009.

Hacıgüzeller, Piraye, James Stuart Taylor, and Sara Perry. 2021. ‘On the Emerging Supremacy of Structured Digital Data in Archaeology: A Preliminary Assessment of Information, Knowledge and Wisdom Left Behind’. *Open Archaeology* 7 (1): 1709–30. https://doi.org/10.1515/opar-2020-0220.

Hayles, N. Katherine. 2012. ‘How We Think; The Digital Humanities’. In *How We Think: Digital Media and Contemporary Technogenesis*, 1–54. Chicago: University of Chicago Press. http://www.press.uchicago.edu/ucp/books/book/chicago/H/bo5437533.html.

Howard-Johnston, James, ed. 2020. *Social Change in Town and Country in Eleventh-Century Byzantium*. Oxford Studies in Byzantium. Oxford, New York: Oxford University Press.

Jeffreys, Michael. 2009. ‘The Seals Module of the Prosopography of the Byzantine World’. *Byzantinoslavica - Revue internationale des Etudes Byzantines* 67 (1–2): 17–23.

Jeffreys, Michael. 2017. *Prosopography of the Byzantine World*, 2016. King’s College London. https://pbw2016.kdl.kcl.ac.uk/.

Kahn, Rebecca, and Rainer Simon. 2020. ‘Feast and Famine: The Problem of Sources for Linked Data Creation’. In *Graph Technologies in the Humanities - Proceedings 2020*, edited by Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera, and Joris van Zundert, 3110:86–100. CEUR Workshop Proceedings. Vienna, Austria: CEUR. https://ceur-ws.org/Vol-3110/#paper5.

Lilie, Ralph-Johannes, Claudia Ludwig, Thomas Pratch, and Ilse Rochow, eds. 2013. *Prosopographie Der Mittelbyzantinischen Zeit Online*. Berlin, Boston: De Gruyter. https://www-degruyter-com.uaccess.univie.ac.at/view/db/pmbz.

Matthew of Edessa. 2017. *The Chronicle of Matthew of Edessa*. Translated by Robert Bedrosian. Long Branch, N.J. http://archive.org/details/ChronicleMatthewEdessa.

McCarty, Willard. 2004. ‘Modeling: A Study in Words and Meanings’. In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 254–70. Oxford: Blackwell. http://www.digitalhumanities.org/companion/.

Mouritsen, Henrik, John Bradley, Dominic Rathbone, and Maggie Robb. 2015. ‘DPRR: Digitizing the Prosopography of the Roman Republic’. In https://dh-abstracts.library.cmu.edu/works/2238.

Nelson, Janet L., and Francesca Tinti. 2006. ‘The Aims and Objects of the Prosopography of Anglo-Saxon England: 1066 and All That?’ In Name Und Gesellschaft Im Frühmittelalter. Personennamen Als Indikatoren Für Sprachliche, Ethnische, Soziale Und Kulturelle Gruppenzugehörigkeiten Ihrer Träger, edited by Dieter Geuenich and Ingo Runde, 241–58. Hildesheim ; Zürich ; New York: Georg Olms Verlag.

Oldman, Dominic, Martin Doerr, and Stefan Gradmann. 2015. ‘Zen and the Art of Linked Data’. In A *New Companion to Digital Humanities*, 251–73. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118680605.ch18.

Orlandi, Tito. 1999. ‘The Scholarly Environment of Humanities Computing’. May 1999. http://web.archive.org/web/20100210151337/http://rmcisadu.let.uniroma1.it/~orlandi/mccarty1.html.

Presner, Todd, Jeffrey Schnapp, and Peter Lunenfeld. 2009. ‘The Digital Humanities Manifesto 2.0’. http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.

Seibt, Werner. 2010. ‘Übernahm der französische Normanne Hervé (Erbebios Phrangopolos) nach der Katastrophe von Mantzikert das Kommando über die verbliebene Ostarmee?’ In *Studies in Byzantine Sigillography* 10, edited by Jean-

Claude Cheynet and Claudia Sode, 89–96. Berlin / New York: Walter de Gruyter. https://www.degruyter.com/document/doi/10.1515/9783110227055.89/html.

Unsworth, John. 2002. ‘What Is Humanities Computing and What Is Not?’ 8 November 2002. http://computerphilologie.uni-muenchen.de/jg02/unsworth.html.

Vogeler, Georg. 2021. ‘“Standing-off Trees and Graphs”: On the Affordance of Technologies for the Assertive Edition’. In edited by Elena Spadini, Francesca Tomasi, and Georg Vogeler, 15:73–94. Norderstedt: BoD. http://www.uni-koeln.de/.

RELAZIONI DEGLI INVITED SPEAKERS

Ricostruzione del testo e banche dati. La filologia digitale alla prova dell'esegesi antica della *Commedia*[1]

Vittorio Celotto, Andrea Mazzucchi
Università di Napoli "Federico II", Italia

L'obiettivo di questo intervento sarà quello di verificare come, in determinate condizioni, la cultura del libro manoscritto e quella dei testi prodotti in era informatica, pur distanti tra loro nel tempo e nelle realizzazioni materiali, tendano ad attuare procedimenti sorprendentemente simili. A fronte della rigidità e stabilità del testo a stampa, la dimensione virtuale, elastica, fluida, instabile del testo elettronico, presenta nel sistema di produzione e circolazione, non meno che nella morfologia testuale, analogie e affinità non trascurabili con la tradizione manoscritta medievale, renitente a ogni forma di standardizzazione e uniformità, che invece, come è noto, è la cifra idiosincratica della stampa.
Si tenterà contestualmente di riflettere sull'insufficienza delle tradizionali procedure filologiche, culminanti nell'edizione critica a stampa, e di richiamare l'attenzione sulle opportunità che l'ambiente digitale e l'edizione elettronica possono offrire per restituire in termini di moderna ordinaria leggibilità le complesse, magmatiche, plurilivellari, instabili, mutevoli testualità di molte opere medievali. Nondimeno si dovranno però valutare attentamente i rischi di un troppo fideistico entusiasmo verso le realizzazioni della cosiddetta filologia digitale che, pur offrendo indiscutibili supporti, non può e non deve sempre costituire l'unico orizzonte entro cui orientare l'attività ecdotica.

1. Nell'ultimo ventennio sono stati descritti con ricchezza di dati, ampiezza di prospettiva e finezza teorica alcuni elementi decisivi per una più corretta descrizione morfologica della testualità manoscritta di epoca medievale. Le accurate diagnosi hanno conseguentemente prodotto più efficaci strategie di resa editoriale e valutazioni ermeneutiche e ricostruzioni storico-letterarie più attendibili e meno anacronistiche. Gran parte delle questioni sono riconducibili a ciò che, con una fortunata quanto efficace definizione, Alberto Vàrvaro ha definito «un testo a costanza narrativa debole» e «a basso gradiente d'autorialità».[2] I testi in volgare, ma anche molte tipologie di testi mediolatini, potevano essere rimaneggiati e modificati, incorporati e assemblati in nuovi organismi testuali, in ragione degli interessi e dei gusti di nuovi lettori. Deriva da ciò una duplice conseguenza:

a) Un'endemica e difficilmente razionalizzabile variabilità delle tradizioni manoscritte, che di fatto stabilisce un *continuum* tra la copia e il rifacimento; si assiste cioè a una costante alterazione dei testi, che perdono spesso la loro fisionomia originaria, perché sottoposti ai filtri pragmatici imposti dall'uso. I testi presentano spesso una struttura che è stata efficacemente definita, sulla scorta di una fortunata formula sociologica di Bauman, "liquida", «perché non rigida e ferma, ma instabile e pronta ad assumere ogni volta la forma, sempre diversa, che decide di fargli assumere lo scriba o, se si spinge sensibilmente oltre il suo compito di semplice copista, il nuovo autore»[3]. Il fenomeno è talmente pervasivo da far dubitare talvolta della pertinenza metodologica di una distinzione tra tradizione diretta e indiretta di un testo, distinzione evidentemente legata ad una concezione moderna dell'opera come entità unitaria e perfettamente circoscrivibile.

b) Un fortissimo nesso fra operazioni testuali e supporti materiali; fra il codice, il libro e il testo, l'opera, con conseguente necessaria attenzione a quella che è stata efficacemente definita la "politica semiotica dei copisti". A partire dagli studi di Domenico De Robertis sui canzonieri si è sviluppata la consapevolezza che un manoscritto miscellaneo non è solo un contenitore di testi, bensì una struttura organica che con maggiore o minore coerenza realizza (o può realizzare) un sistema semiotico produttore di senso. Altrettanto rilevanti in questa direzione sono gli studi di Keith Busby, per i quali l'identità stessa di un testo si definisce a partire dal suo inserimento in una determinata raccolta manoscritta.[4] Testi e libri, soprattutto se privi di una marcata dimensione autoriale e di una rigida strutturazione formale, subiscono infatti frequentissime operazioni di frammentazione e di nuovi accorpamenti, con espunzioni di sezioni, assemblaggi di nuove

[1] Il testo riproduce l'intervento tenuto in occasione del XIII Convegno AIUCD 2024 (Catania, 28-30 maggio). Si devono ad Andrea Mazzucchi i parr. 1-4, a Vittorio Celotto i parr. 5-8, ma s'intende che il contributo è stato concepito ed elaborato di concerto dai due autori.
[2] Cfr. A. Vàrvaro, *Il testo letterario*, in *Lo spazio letterario del Medioevo. 2. Il Medioevo volgare*, vol. I *La produzione del testo*, to. I, a cura di P. Boitani, M. Mancini e A. Vàrvaro, Roma, Salerno Editrice, 1999, pp. 387-422.
[3] F. Delle Donne, *Testi "liquidi" e tradizioni "attive" nella letteratura cronachistica mediolatina*, in *Il testo nel mondo greco e latino*, a cura di G. Polara e A. Prenner, Napoli, Liguori, 2015, pp. 15-38.
[4] Cfr. K. Busby, *Codex and Context. Reading Old French Verse Narrative in Manuscripts*, Amsterdam-New York, Faux Titre, 2002.

porzioni, fusioni e contaminazioni, che danno vita a costellazioni testuali dalla fisionomia culturale sempre diversa. Tipico il caso dei testi esegetici, di quelli agiografici, delle cronache, della letteratura didattica in prosa.
Il ruolo fortemente attivo della tradizione si esplica non solo nella selezione dei testi e nel loro trattamento redazionale, ma come è ovvio anche nella loro *mise en page* e nell'eventuale corredo paratestuale. È divenuto sempre più decisivo analizzare il rilievo del testimone non solo in termini filologici e storico-tradizionali (estrazione socio-culturale del copista, usi grafici, connotati linguistici), ma di considerare tali indicazioni testuali in maniera solidale con la *mise en texte* del manoscritto, la sua poetica visiva (Wayne Storey), la strategia paratestuale che esso sottende. Di qui il rilievo assunto dalla cosiddetta filologia materiale, anche in Italia, tradizionale roccaforte di procedure ecdotiche su basi genealogico-stemmatiche, volte più che all'analisi del singolo testimone alla ricostruzione di un testo. Appare ormai ineludibile che una buona edizione – e non solo per i testi medievali – tenga presente, per quanto possibile, elementi extratestuali e paratestuali strettamente connessi al testo e imprescindibili per restituirne compiutamente la dimensione di senso.

2. Per illustrare con casi concreti alcuni dei problemi a cui si è accennato e per chiarire con più evidenza le questioni poste in termini generali mi servirò di un esempio desunto dalla tradizione esegetica antica sulla *Commedia*, ma avrei potuto ricorrere abbastanza facilmente ad altre tradizioni retorico-discorsive, avvertendo sin d'ora che per ragioni di spazio e di tempo dovrò però sorvolare su molti dettagli.
Si consideri il caso di molti sistemi irrelati di glosse al poema dantesco, depositatesi sui vivagni dei manoscritti tre-quattrocenteschi o sugli incunaboli della *Commedia*, spesso in intrigante quanto complesso dialogo con vignette e illustrazioni che nella colonna di scrittura o sui margini impreziosiscono quei libri. Per evitare pronunciamenti troppo generali e per verificare su di un caso concreto le questioni che si intendono proporre, partirei dall'analisi delle cosiddette *Chiose filippine* alla *Commedia* tràdite dal solo manoscritto CF 2 16 della Biblioteca Oratoriana dei Girolamini di Napoli. La loro edizione dunque si identifica di fatto con la resa del *codex unicus*, il quale, e qui sta per noi il punto, si configura come un affascinante organismo, organizzato quasi come in un complesso congegno multimediale.[5] Non solo sul manoscritto si avvicendano ben sei differenti mani operanti tra la metà del XIV e la fine del XV secolo, ma provvedono a facilitare la fruizione e la memorizzazione del testo dantesco anche 146 vignette miniate, di cui 131 inserite nella colonna di scrittura e 15 sui margini delle carte. E in aggiunta numerose postille marginali alle vignette illustrative, sorta di rubriche ausiliarie, di sintetiche didascalie, che ritraducono verbalmente il contenuto dell'illustrazione, provvedendo verosimilmente a precisare la porzione del testo dantesco cui la vignetta pertiene.
Nel caso della miniatura a c. 60*v*, che illustra l'incontro di *Inf*., XXV con il centauro Caco, la chiosa «Iste debuit pingi ad modum centauri» vale, per esempio, nelle intenzioni del chiosatore, a correggere una svista dell'illustratore effettivamente riscontrabile nella vignetta relativa.
A complicare il congegno testuale, si registrano poi frequenti interventi di modifica e integrazione sul testo del primo copista, che nelle intenzioni del nuovo lettore migliorano e correggono il testo di partenza, ma che determinerebbero incoerenze testuali significative, se li si riproducesse nella loro illusoria uniforme linearità, cancellandone, per così dire, la pluridimensionale verticalità diacronica. Si considerino, tra i tanti, il caso della postilla a *Inf*., I 101 sul veltro, che il primo anonimo glossatore identifica con «quidam dominus temporalis», cui segue la proposta di un secondo più tardo postillatore che aggiunge «et religiosus et sic videtur pronosticare de veltro». Letta nella sua illusoria continuità sintagmatica, senza cioè artifici grafici che consentano di differenziare in diacronia le porzioni di testo, la chiosa restituisce al lettore della moderna edizione un ircocervo interpretativo intimamemente contraddittorio, un insanabile ossimoro esegetico.
Analogamente Mondret, nella chiosa a *Inf*., XXXII 61, è riconosciuto dal primo postillatore, forse sulla scorta di Geoffrey di Monmouth, come «nepos regis Artus». Ma su questa pericope testuale interviene, in una fase successiva, la modifica del secondo postillatore, che nell'interlinea inserisce, tra *nepos* e *regis*, «imo filius», con l'evidente intenzione di migliorare il testo, correggendolo con un'informazione nuova ma potenzialmente contraddittoria con la precedente.
La pagina straordinariamente affascinante del Filippino induce dunque subito nel lettore-osservatore inquietanti e allarmanti interrogativi sulla possibilità di fornire, in termini editoriali moderni, nel pur complesso congegno di un'edizione critica a stampa, non solo un plausibile corrispettivo, ma anche un resoconto esauriente delle varie fenomenologie. L'esigenza ovvia di rendere il testo leggibile in tutte le sue complicate articolazioni pare scontrarsi senza rimedio con la necessità, altrettanto ineludibile, di render conto della forma del testo, essenziale per una ragionevole decodifica del materiale esegetico adibito.
La natura stessa del manoscritto ne ha imposto, dunque, in sede di edizione, una resa integrale, tale da render conto della complessa stratificazione, sia sostanziale che formale, da cui il libro sorprendentemente trae origine. Ci troviamo infatti di

[5] Cfr. *Chiose filippine. Ms. CF 2 16 della Bibl. Oratoriana dei Girolamini di Napoli*, a cura di A. Mazzucchi, Roma, Salerno Editrice, 2002.

fronte a chiose irrelate e non ad un commento organico, e la dislocazione dei vari elementi sulla pagina è fattore fondamentale per una corretta lettura. La sola rappresentazione del materiale verbale nell'edizione ha dovuto prevedere una serie di accorgimenti tecnico-tipografici (devo dire fortunati) non usuali che, come d'obbligo in un'edizione critica, ma in questo caso con un evidente surplus di necessità e insieme di difficoltà, hanno tentato di fornire al fruitore tutti gli strumenti necessari a una sorta di ardua proiezione ricostruttiva della morfologia materiale della pagina. In un organismo così strategicamente congegnato e insieme così stratificato in diaconia come questo del Filippino, i rapporti sulla pagina si caricano di un potenziale esegetico molto cospicuo, cui non è lecito rinunciare se si vogliono evitare letture parziali o equivoche.
L'edizione cartacea ha però potuto risolvere solo in parte la questione del commento figurato. Si è infatti fornita una sintetica descrizione del contenuto iconografico di ogni vignetta, ogniqualvolta è stato necessario riportare quelle chiose marginali alle miniature di cui si è parlato. Si tratta, come è facile comprendere, di una misura minima, del tutto inadeguata al rilievo inoppugnabile che la serie illustrativa assume nel manoscritto.

3. Le stratificate articolazioni, con aggiunzioni progressive su un impianto testuale già definito, il complesso gioco di rimandi e richiami tra il testo dantesco primario, le illustrazioni che ne enfatizzano alcuni episodi e i brandelli testuali prodottisi a partire dalle illustrazioni, che rendono così affascinanti le carte del ms. Filippino, forniscono una plastica conferma all'ipotesi di chi ha recentemente avvicinato la testualità medievali ad uno dei più fortunati fenomeni della cultura contemporanea: Wikipedia. Anche l'anonimo redattore di Wikipedia, come i diversi anonimi menanti che si sono avvicendati sulle carte e nei margini del ms. Filippino, può liberamente riscrivere, rimaneggiare la voce scritta da un precedente, ma altrettanto anonimo, redattore:

> Se un qualsiasi copista poteva prendersi la libertà di riscrivere o aggiornare parti più o meno ampie del *Trésor* di Brunetto Latini, tanto più (e talora con meno competenze) un redattore anonimo di Wikipedia può rimaneggiare interi periodi di una voce scritta da un altrettanto sconosciuto autore.[6]

Il testo è il risultato di una stratificazione d'interventi disomogenei e diacronicamente successivi potenzialmente mai interrotti. Né, come è stato giustamente osservato, le analogie terminano qui. Bisognerà considerare anche il basso gradiente di autorialità delle voci di Wikipedia, che si prestano, al pari molte opere medievali, a modifiche, riscritture, selezioni, tagli, interpolazioni e contaminazioni dei diversi utenti; la struttura fortemente entropica dell'accumulo di materiali che spesso approda alla inevitabile necessità di radicali riformulazioni finalizzate a restituire una qualche forma di leggibilità alle voci; la gerarchizzazione e la tassonomia delle informazioni improntata non alle sequenze sintagmatiche dell'ordine alfabetico tipiche delle enciclopedie a stampa, ma piuttosto ad associazioni paradigmatiche non distanti dai meccanismi paraetimologici della tradizione lessicografica medievale.
E tuttavia le pur impressionanti affinità fin qui sottolineate tra la testualità medievale e quella dell'era informatica non intendono affatto suggerire una riproposta – davvero fuori tempo massimo, soprattutto dopo le lucidissime prese di distanza di Alberto Vàrvaro – delle confuse posizioni post-strutturaliste, barthesiane e foucaultiane della *new philology*, assunte quale fondante archetipo metodologico e rilevante giustificazione storico-teorica delle più ardite pratiche di filologia digitale.[7] Nella ricostruzione di Cerquiglini i vincoli imposti dalla carta stampata non consentono una autentica restituzione della fluidità dei testi medievali. L'edizione critica tradizionale di tipo genealogico-ricostruttivo, nel suo illusorio tentativo di razionalizzare una tradizione fortemente attiva, nello sforzo di ricondurre a un'unità teorica e astratta la molteplicità delle singole concrete testimonianze, non sarebbe in grado di restituire ai lettori di oggi l'autentica storicità del testo medievale, intrinsecamente metamorfico, pluriforme e renitente a ogni forma di moderna convenzionale costrizione (Cerquiglini 1989, ma anche Frank 1993).
Le edizioni critiche, in particolare, data la loro complessità, potrebbero trarre grandi vantaggi dall'informatica, perché essa potrebbe permettere di superare i limiti del supporto cartaceo, cioè la bidimensionalità. «Un'edizione critica, anche nella sua forma più tradizionale, è in un certo senso un'opera tridimensionale: il testo si sviluppa sulla dimensione piana della pagina, ma l'apparato critico costituisce una sorta di approfondimento verticale, una terza dimensione del testo», quella della diacronia.[8]
Tanto nella digitalizzazione delle testimonianze quanto nella resa della stratificazione diacronica dei testi (varianti d'autore o interventi successivi di altre mani), il supporto elettronico ha moltiplicato le possibilità dell'edizione critica. L'idea di

[6] Cfr. C. Lagomarsini, *Wikipedia e la "tradizione aperta"*, in www.claudiogiunta.it.
[7] Cfr. in particolare A. Vàrvaro, *La New Philology nella prospettiva italiana*, in *Alte und neue Philologie*, a cura di M-D. Glessgen e F. Lesbanft, Tübingen, Niemeyer, 2007, pp. 35-42.
[8] Cfr. P. Chiesa, *Elementi di critica testuale*, Bologna, Patron, 2012.

testo stabile, che fa da sfondo alla filologia scientifica del XIX secolo, è un'idea interna alla civiltà della stampa: la civiltà del manoscritto – lo si è detto – spesso non riconosceva l'idea di stabilità del testo. Ora i testi digitali sembrerebbero possedere tutte le caratteristiche per offrire ambienti capaci di tradurre in termini di ordinaria leggibilità i complessi organismi testuali medievali.
A differenza di un'edizione critica a stampa, l'edizione digitale parrebbe configurarsi infatti come un oggetto dinamico, non statico. Un perfetto strumento per rendere la *mouvance* del testo. Non solo. Esso offre infatti anche la possibilità di gestire più piani allo stesso tempo, garantendo un raccordo assai più stretto e un efficace trasferimento del rapporto testo-immagine e restituendo le varie morfologie paratestuali. Grazie alla tecnologia recuperiamo altre dimensioni alla narrazione verbale che pure erano previste nella cultura del libro manoscritto. Il formato, l'aspetto fisico e materiale del codice erano elementi capaci di condizionare la concezione del testo. Ne resta ancora una qualche traccia nel capolavoro assoluto della letteratura medievale, in un'opera pur segnata da un'autorialità forte e orchestrata su una rigida quanto mirabile programmazione testuale. Gli ultimi versi del *Purgatorio* dantesco infatti alludono – certo solo per un'efficace strategia comunicativa – a una conclusione della cantica imposta da ragioni strettamente materiali: l'esaurimento dello spazio delle carte del fascicolo predisposte a ospitare il testo: «S'io avessi, lettor, più lungo spazio / da scrivere, i' pur canterei in parte / lo dolce ber che mai non m'avria sazio; / ma perché piene son tutte le carte / ordite a questa cantica seconda, / non mi lascia più ir lo fren dell'arte» (*Purg*., XXXIII 136-41).
L'edizione critica digitale sembrerebbe dunque la soluzione ideale per offrire ai lettori contemporanei un'immagine storicamente fededegna e filologicamente rispettosa di molte testualità medievali, laddove la rigida linearità e la forzata unicità dell'edizione critica a stampa non sarebbero capaci di restituire, pur nella proiezione stratigrafica di testo e apparati, la consustanziale molteplicità di esecuzioni scritte e orali delle letterature romanze e mediolatine.

4. Ragionare però in termini di contrapposizioni polari rischia di semplificare un quadro che è assai più mosso e complicato. Si dovranno infatti valutare, prima di aderire fideisticamente alle novità dell'edizione critica digitale, non solo alcuni limiti, già più volte segnalati nella bibliografia sul tema, quali l'assenza di criteri di valutazione riconoscibili e condivisi per le risorse digitali simili a quelli messi in atto per valutare un libro a stampa tradizionale; la costante cura e manutenzione che un'edizione digitale richiede per continuare ad essere accessibile; la necessità di competenze tecniche elevate, garantite solo da gruppi di lavoro con esperti provenienti da diverse discipline; la difficile valutazione dell'apporto individuale nel modello collaborativo che sta alla base delle creazioni di risorse digitali; l'affidabilità – cruciale nelle discipline umanistiche – nella citazione di prodotti per loro natura predisposti alla modifica e che si prestano potenzialmente a costanti aggiornamenti. Se è ragionevole ipotizzare che lo sviluppo e l'affermarsi delle *digital humanties* renderà meno cogenti tali limiti, come in parte già sta accadendo, più radicale è l'interrogativo che investe i criteri filologici che tali nuove pratiche editoriali presuppongono o posso favorire. Bisognerà infatti a tale proposito distinguere con attenzione in relazione almeno a tipologie testuali e a morfologie delle tradizioni.
L'edizione elettronica, con la possibilità di affiancare la riproduzione digitale del manoscritto alla restituzione in termini di moderna e ordinaria leggibilità del testo che vi è contenuto, sembra infatti costituire la soluzione ideale per tradizioni monotestimoniate e soprattutto per opere che non sono state concepite per la pubblicazione né per la circolazione al di fuori di una cerchia personale o familiare, come nel caso dei testi esegetici nati per uso privato, o nel caso delle cronache, dei libri di famiglia, delle scritture memorialistiche, spesso configurantesi come esempi di *work in progress*, dove scrittura e lettura si svolgono intorno a un singolo manufatto, lasciando tracce e stratificazioni caratteristiche.
Ma per tipologie testuali a più forte gradiente di autorialità, con opere rigorosamente improntate ai criteri di coerenza e coesione testuale e soprattutto per testi a tradizione plurima privi di autografo, l'edizione critica digitale non solo ancora non costituisce l'orizzonte prevalente entro cui muoversi, ma può determinare pericolose semplificazioni e gravi banalizzazioni metodologiche. In questi casi infatti l'edizione elettronica si è spesso trasformata, anche nelle realizzazioni più alte, solo in un utilissimo, prezioso archivio delle testimonianze, di cui si è rinunciato a ogni tentativo di razionalizzazione e organizzazione gerarchica, limitandosi a giustapporre in modo neutro le diverse realizzazioni di un testo. Come ha lucidamente scritto Lino Leonardi, «nella maggior parte delle edizioni digitali, anche quelle che non si limitano a dar conto di un solo testimone, l'attenzione è focalizzata molto più sulla riproduzione, diciamo pure sull'edizione, di ciascun singolo manoscritto, e sulla corrispondenza di questa alla pagina visualizzata del codice, o al massimo alla possibilità di affiancare i diversi individui nelle finestre dello schermo, che non sulle potenzialità di un confronto approfondito sul piano testuale tra le diverse unità testimoniali».[9]

[9] Cfr. L. Leonardi, *Filologia elettronica tra conservazione e ricostruzione*, in *Digital Philology and Medieval Texts*, ed. by A. Ciula and F. Stella, Pisa, Pacini, 2007, pp. 65-75.

L'appiattimento sulla sincronia delle singole testimonianze, al limite sulla multipla sincronia delle diverse testimonianze come pure il frequente ricorso a edizioni a stampa realizzate secondo il metodo della copia scribale, non solo impediscono di cogliere la diacronia dei processi di una tradizione testuale, di analizzare il diasistema costituito dall'interazione di un testimone con i suoi modelli, ma, ben lungi dal restituire una concreta storicità, di fatto escludono anche la possibilità di valutare la pertinenza culturale di ogni singola testimonianza. Pertinenza che si può misurare solo in termini differenziali e comparativi con le altre testimonianze e con un'ipotesi ricostruttiva, quale solo un'accurata, non rinunciataria, "tradizionale" *recensio* è in grado di offrire.
Non si tratta dunque di opporre e di cristallizzare assiologie tra la vecchia e la nuova filologia, tra le edizioni cartacee a stampa e quelle digitali ed elettroniche, ma di valutare piuttosto di volta in volta il mezzo più adeguato a restituire una sempre più penetrante interpretazione del testo e della tradizione che lo testimonia. Senza neppure escludere e anzi favorendo le soluzioni miste, in cui accanto alla verità del documentato si affianchi la verità del ricostruito, accanto alla sincronia del dato la diacronia del processo, accanto a un testo a stampa, rappresentato in forma continua, riproduzioni virtuali dell'opera nella sua forma fisica originaria o presunta tale.[10]

5. Si intende adesso spostare leggermente il punto di vista dal campo dell'edizione critica digitale a quello delle risorse informatiche utili al potenziamento delle ricerche linguistiche, filologiche e letterarie sull'esegesi dantesca antica, illustrando sinteticamente un progetto di umanistica digitale in corso di realizzazione, grazie a al finanziamento per Progetti di Ricerca di Rilevante Interesse Nazionale del 2022, [11] e che rientra nell'ambito di un più ampio e articolato sistema di risorse digitali sulle opere di Dante e sulla loro tradizione e ricezione già in parte realizzato, e in costante implementazione, presso il Dipartimento di Studi Umanistici dell'Università degli Studi di Napoli Federico II, per la cui descrizione si rinvia al sito www.dante.unina.it.
Il progetto CoDA (Commenti Danteschi Antichi) mira all'elaborazione di una banca dati digitale dei commenti antichi alla *Commedia*, prodotti sia in volgare sia in latino entro la fine del XV secolo. A più riprese auspicata dalla comunità scientifica, essa costituirà un corpus autonomo all'interno del corpus OVI, la raccolta online più ampia e affidabile di testi italiani antichi, che costituisce la base per la compilazione del Tesoro della Lingua Italiana delle Origini (TLIO), vocabolario delle varietà italoromanze delle Origini.
Il progetto prevede:

1) Elaborazione della Bibliografia dei commenti danteschi consultabile online mediante il software PLUTO;
2) Creazione del corpus e sua integrazione nel corpus OVI;
3) Lemmatizzazione dei testi ed elaborazione del sistema di interrogazione secondo i parametri del software GATTO 3.3 e GattoWeb già ampiamente in uso per tutti i corpora OVI;
4) Estensione del software all'interrogazione finalizzata a ricerche di interesse esegetico, relative alla spiegazione del dettato dantesco e ai diversi contenuti dei commenti.

Come il TLIO e il Corpus OVI, il corpus CoDA sarà ad accesso libero e gratuito con la garanzia di una piena sostenibilità grazie alla stabilità della struttura OVI-CNR, depositaria per statuto di una missione di lungo periodo.

6. Il corpus CoDA è una banca dati sviluppata nel software Gatto 3.3 e GattoWeb che raccoglierà tutti i testi dei commenti alla *Commedia* in volgare e in latino databili entro il XV secolo. Si tratta di una delle prime continuazioni del corpus OVI oltre il limite del XIV secolo e della prima integrazione di testi latini interrogabili contestualmente a quelli in volgare.
La finalità del corpus è duplice: lessicografica ed ermeneutica. Il primo obiettivo riguarda lo studio del lessico dei commenti, che contribuisce all'arricchimento del profilo linguistico dell'Italia delle Origini e del quadro dei rapporti tra latino e volgare nelle scritture di tipo espositivo e argomentativo. Il secondo obiettivo consiste nell'incentivare l'analisi comparativa dei commenti rispetto alle prospettive esegetiche sul poema e ai più vasti contenuti culturali che i testi restituiscono.
I testi che confluiranno nel corpus CODA si possono ascrivere a un'ampia gamma di tipologie, ma sono accomunati dall'intento di offrire un ausilio interpretativo al poema dantesco. Si tratta di commenti organici e continui a una cantica o all'intera opera (es. il commento latino al solo *Inferno* del notaio bolognese Graziolo Bambaglioli e i suoi due volgarizzamenti, oppure il primo commento fiorentino integrale al poema noto con il nome di *Ottimo commento*, entrambi compilati tra gli anni '20 e '30 del XIV secolo, a ridosso della morte del poeta); sistemi compositi di postille illustrative

[10] Per un recente ed equilibrato bilancio sull'ormai indispensabile, ma nondimeno dialettico, rapporto tra i criteri metodologici della filologia ricostruttiva di stampo lachmanniano-maasiano e le fruttuose innovazioni della filologia digitale si vd. L. Leonardi, *Filologia digitale del medioevo italiano*, in «Griseldaonline», 20 2, 2021, pp. 77-89.
[11] Il progetto, da me coordinato, si avvale della partecipazione delle Unità di ricerca locali dell'Istituto dell'Opera del Vocabolario Italiano (Zeno Verlato) e di Sapienza Università di Roma (Luca Fiorentini).

(es. le cosiddette *Chiose Palatine* del ms. Laur. Pal. 313 in volgare fiorentino, o le cosiddette *Chiose Filippine* in latino, con inserti in volgare napoletano, del ms. CF 2 16 della Biblioteca Oratoriana dei Girolamini di Napoli, stratificatesi nel corso di più di un secolo, dal 1360 ca. alla fine del '400); lezioni pubbliche (es. le *recollectae* dei corsi tenuti da Benvenuto da Imola a Bologna e a Ferrara nel biennio 1375-1376); compendi in prosa o in versi (es. la *Divisione* del figlio di Dante, Iacopo Alighieri, riassunto in terzine della struttura del poema); traduzioni (es. la prima traduzione latina del poema, stesa da Giovanni Bertoldi da Serravalle nel 1416 in occasione del Concilio di Costanza).
Alcuni testi, anche quando anonimi, sono caratterizzati da un forte carattere autoriale, che si rivela in una prospettiva ideologica riconoscibile; altri sono piuttosto complessi apparati notulari, che restituiscono inediti spunti esegetici, ma mancano di una struttura organica e di una concezione unitaria.
I testi sono inoltre caratterizzati da un forte ibridismo linguistico e culturale, dovuto al fatto che gli autori esibiscono profili professionali e intellettuali diversi: scrittori di raffinatissima cultura letteraria come Giovanni Boccaccio, protagonisti della vita di corte come Guglielmo Maramauro, funzionario della regina Giovanna I di Napoli; personaggi afferenti all'ambiente notarile e giuridico come Andrea Lancia, ecclesiastico come Guido da Pisa, universitario come Francesco da Buti.
A partire da un'attenta ricognizione che ha incrociato i dati ricavabili dai repertori e dalle ultime ricognizioni sullo stato degli studi sono stati individuati circa 40 testi, tra commenti e sistemi di chiose, che sono i commenti di tradizione manoscritta fino al 1478-1480, cioè fino alla pubblicazione dei commenti di Nidobeato e Cristoforo Landino.[12] Nel corpus sono naturalmente inclusi anche i volgarizzamenti di testi latini circolanti entro quella data.
Il limite cronologico dipende da considerazioni di carattere storico, giacché a partire dall'apparizione dei primi commenti stesi appositamente per la stampa, lo statuto del genere muta sensibilmente. I testi di tradizione manoscritta implicano infatti un più sfumato concetto di autorialità, con conseguenze importanti sia sul piano filologico e linguistico sia su quello della loro valutazione critica. La fine del Quattrocento costituisce peraltro uno spartiacque importante nella critica dantesca tra una fase ancora legata al commento medievale a carattere enciclopedico e una che guarda al poema con interesse esclusivamente letterario; infine, i commentatori più antichi condividono con Dante lo stesso orizzonte culturale e le stesse passioni ideologiche, mentre in epoca umanistica la distanza da quell'orizzonte lascia il posto a un approccio antiquario alla *Commedia*.
In quanto integrazione del corpus OVI, il corpus CoDA ne erediterà il metodo di raccolta e i punti di forza: le molteplici potenzialità di interrogazione del software GattoWeb (attraverso la lemmatizzazione esaustiva dei testi), l'associazione del testo ai passi del poema corrispondenti, l'associazione dei volgarizzamenti al testo di traduzione, l'integrazione di una bibliografia completa in PLUTO che consentirà di collegare i testi con informazioni sulla loro tradizione e vicenda editoriale.

7. Da questo punto di vista, il vantaggio del corpus CoDA sugli strumenti già esistenti dedicati ai commenti danteschi, è che, in virtù della collaborazione tra OVI-CNR e Centro Pio Rajna, potrà contare per la prima volta sulla digitalizzazione integrale di tutti i testi che sono stati pubblicati (o che sono in via di pubblicazione) nell'ambito della «Edizione Nazionale dei Commenti Danteschi». Il corpus CoDA sarà dunque la prima raccolta digitale online che potrà avvantaggiarsi delle edizioni condotte con attendibili criteri filologici e linguistici che hanno caratterizzato la più recente stagione degli studi danteschi. Per i testi di cui non è ancora disponibile un'edizione recente, i criteri di costituzione del corpus CoDA si integrano con la linea operativa del corpus OVI, accogliendo le edizioni disponibili, anche quando datate, che possono comunque essere utilmente considerate in una prospettiva che ne recuperi i soli dati lessicali. A seguito di opportuni controlli, verrà introdotta infatti la distinzione (già in uso nel corpus OVI) tra testi significativi per la descrizione linguistica di una determinata varietà ("testi TS") e testi meno attendibili e che abbisognano di ulteriori verifiche.
Si fornisce un esempio dei vantaggi di questo genere di acquisizione. Finora il commento di Iacomo della Lana poteva essere consultato e interrogato esclusivamente in un'edizione ottocentesca (che è quella presente in tutti gli strumenti digitali a disposizione), caratterizzata da un massiccio interventismo sia sui fatti di sostanza (tagli, interpolazioni, spostamenti ecc.) sia sulla patina linguistica, ortopedizzata in direzione fiorentina. La recente edizione a cura di Mirko Volpi presenta invece una versione sinottica dei due principali rami linguistici della tradizione, che consente di leggere il testo del ms. Riccardiano-Braidense, che tra i testimoni conservati restituisce la veste linguistica bolognese più prossima all'originale (oltre a essere naturalmente un raro monumento di scrittura bolognese trecentesca), e quello del ms.

[12] Cfr. S. Bellomo, *Dizionario dei commentatori danteschi. L'esegesi della 'Commedia' da Iacopo Alighieri a Nidobeato*, Firenze, Olschki, 2004; *Censimento dei commenti danteschi. 1. I commenti di tradizione manoscritta (fino al 1480)*, a cura di E. Malato e A. Mazzucchi, Roma, Salerno Editrice, 2011; A. Mazzucchi, *Questioni di metodo sull'edizione degli antichi commenti alla 'Commedia'*, in «Rivista di Studi Danteschi», a. XVIII fasc. 1 2018, pp. 153-71.

Trivulziano 2263, giudicato il più autorevole della fortunata tradizione toscana.[13] Si comprende immediatamente l'importanza che l'immissione nel corpus CoDA della nuova edizione fornisce in prospettiva tanto testuale quanto lessicografica.

8. Una piattaforma digitale come il corpus CoDA risponde all'esigenza di far transitare testi e conoscenze in uno spazio virtuale liberamente accessibile e interrogabile sulla base di parametri chiari e funzionali, così da rendere disponibile un'ingente quantità di dati, facilitandone la consultazione. Ma il progetto non si limita alla costituzione di un corpus: l'aggregazione di testi in banche dati digitali è oggi operazione relativamente semplice, ma rischia di restare una mera operazione quantitativa se non fornisce strumenti e condizioni perché quei dati siano connessi e interrogati opportunamente, offrendo così prospettive particolari e panoramiche generali che possano essere sottoposte all'interpretazione storica.
Alla base del progetto CoDA vi sono tre indirizzi metodologici fondamentali: 1. Il circolo virtuoso tra filologia e lessicografia digitali è costitutivo del progetto, giacché la filologia fornisce testi attendibili su cui effettuare spogli lessicali, mentre la lessicografia fornisce la piena conoscenza delle parole, delle forme e delle costruzioni vigenti nell'età e nell'ambiente a cui un testo appartiene; 2. L'importanza dello studio comparativo del lessico dei commenti non solo finalizzato a ricavare informazioni su una parola, ma ad aprire la strada all'intero sistema della lingua, alle stratificazioni culturali e sociali che la storia semantica delle parole porta con sé, ampliando le nostre conoscenze sulla circolazione dei saperi in diacronia e in sincronia; 3. La necessità di studiare i commenti danteschi sia valorizzandone le differenze (i profili degli autori, le destinazioni e i contesti di produzione di questi testi), sia attraverso uno sguardo d'insieme che vada oltre l'individualità del singolo commento per illuminare la riflessione comune sulla lingua poetica di Dante e sulle sue implicazioni in molteplici campi del sapere medievale.
La legittimità dell'esistenza di un corpus settoriale come CoDA è data da una visione storiografica che giudica gli antichi commenti come un filone di particolare rilievo tra le scritture in volgare e in latino del Medioevo. I tratti peculiari di questo filone si riconoscono nel suo carattere ibrido all'incrocio tra il genere del commento letterario e quello della summa enciclopedica, disponibile cioè ad accogliere generi discorsivi diversi: dalla filosofia alla teologia, dalle scienze naturali all'astrologia, dalla storiografia alla narrativa. Giacché per essere compreso appieno il poema dantesco richiedeva ai suoi lettori competenze culturali di ordine vario e diverso, la necessità di rivelarne il senso a un pubblico ampio, fatto di dotti e *illitterati*, comporta l'approfondimento di tutte le sue implicazioni storiche e dottrinarie.
Il frequente ricorso a materiale testuale allotrio, assemblato attraverso diverse tecniche compositive, dalla parafrasi al rimaneggiamento al volgarizzamento di brani latini, comporta la commistione di filiere culturali differenti e fa di tali apparati esegetici ricchissimi repertori di tipi lessicali associati a campi semantici eterogenei, afferenti alla cultura materiale tanto quanto al linguaggio scritturale, al gergo scientifico o a quello liturgico e così via. Attenti in primo luogo alla spiegazione della lettera del poema, i commenti sono un vastissimo serbatoio di varianti sinonimiche delle parole di Dante: sono particolarmente sensibili alle voci desuete o di nuovo conio della *Commedia* e al suo plurilinguismo espressivo, di cui non di rado forniscono proposte interpretative di ordine diatopico o diastratico; si mostrano attentissimi ai fatti lessicali, non solo per il significato, ma anche per la qualità e la provenienza; offrono numerosi casi di proliferazione di geosinonimi, cortocircuiti onomasiologici, diffrazioni tanto geolinguistiche quanto semantiche e interpretative.

9. Ereditando i punti di forza della Bibliografia dei commenti danteschi e dell'immissione del corpus nel software GattoWeb, il sistema di interrogazione che si intende elaborare punta a superarli in funzione di ricerche di interesse contenutistico e specificamente esegetico.
Difficile sopravvalutare l'utilità dei commenti antichi per la storia della critica dantesca. Essi sono infatti ricchissimi di informazioni storiche e aneddotiche, ipotesi interpretative, notazioni linguistiche, stilistiche e retoriche, collegamenti intertestuali e riflessioni sulle fonti di Dante. Si tratta di un materiale inestimabile, che ci consegna le reazioni dei primi lettori a cui Dante si rivolgeva e che con lui condividono mentalità e cultura.
La storia di un classico come la *Commedia* reca con sé anche la storia della sua tradizione interpretativa. Per questo motivo, gli studi hanno ampiamente rivelato come l'analisi comparativa di commenti di tipologia e provenienza diversa si riveli efficace sia per una più accorta interpretazione di aspetti diversi del testo, sia per illuminare l'orizzonte culturale in cui esso circolava. Questo approccio impone di integrare la considerazione della individualità del singolo commento con una visione d'insieme di tipo comparativo: il confronto di prospettive interpretative molteplici consente infatti di far affiorare una riflessione collettiva e spesso collaborativa intorno alle complesse questioni sollevate dal testo commentato.

[13] Cfr. Iacomo della Lana, *Commento alla 'Commedia'*, a cura di M. Volpi, con la collaborazione di A. Terzi, Roma, Salerno Editrice, 2009.

Il progetto CoDA prevede l'applicazione di un opportuno sistema di marcatura e indicizzazione che consenta di interpellare tutti i testi o una selezione di essi con uno scopo di ricerca comparativo. Si prevede la possibilità di utilizzare parametri di ricerca diversificati e caratterizzati da un elevato livello di sofisticatezza, grazie allo sfruttamento della lemmatizzazione esaustiva operato in GattoWeb. La ricerca sarà eseguibile mediante una *query* disposta ad accogliere singole forme e lemmi oppure cooccorrenze; potrà essere articolata nel modulo della "ricerca esatta" grazie al quale il sistema troverà le forme che corrispondono esattamente alla stringa digitata, oppure, grazie all'utilizzo di caratteri speciali, nel modulo della "ricerca espansa", grazie al quale il sistema trova le forme che contengono la stringa specificata.
Il corpus sarà interrogabile a partire dai seguenti criteri selettivi, che qui si riassumono accompagnandoli con esempi esposti in forma sintetica:

1. Voci della *Commedia*. Il sistema raccoglie i brani con funzione di glossa in cui viene fornita la definizione della voce ricercata. Es. la voce *meare* ('passare, derivare') a *Par*., xiii 55 è interpretata in maniere diverse dall'*Ottimo commento* («cioè che ssi indea Patre – cioè ch'è uno Idio col Padre»); da Francesco da Buti («cioè per sì fatto modo si deriva per generazione»); da Benvenuto da Imola («idest quae unitur et fit ea»).
2. Voci dei commenti. Il sistema raccoglie i brani di diversi commenti in cui compare una determinata forma. Es. la ricerca della voce verbale *metaforizzare* rivela l'uso esclusivo dei più antichi commenti trecenteschi (il bolognese Iacomo della Lana, a cui fanno seguito i fiorentini *Ottimo commento*, Andrea Lancia e Anonimo fiorentino). Si registra una interessante distribuzione che, ad esempio nell'*Ottimo*, vede nessuna occorrenza nelle chiose all'*Inferno*, 1 nel *Purgatorio* e ben 7 nel *Paradiso*, spesso peraltro con accezioni leggermente diverse, a testimonianza di un uso del tecnicismo retorico esteso a figure diverse di figurazione polisemica. Infine, la ricerca integrata nel corpus OVI consente di verificare che il verbo occorre nelle varietà italoromanze del Trecento esclusivamente nei commenti alla *Commedia*.
3. Passi della *Commedia*. Il sistema raccoglie i brani che hanno per oggetto un verso, una terzina o un gruppo di versi più esteso. Ad es. il confronto delle chiose al primo verso del poema («Nel mezzo del cammin di nostra vita») mette in luce l'accordo di quasi tutti i commentatori sul fatto che l'indicazione si riferisca all'età di trentacinque anni, con poche eccezioni. Guido da Pisa chiosa «Medium namque vite humane, secundum Aristotelem, somnus est», coerentemente con la sua lettura della *Commedia* come visione profetica ricevuta in sogno; Filippo Villani all'inizio del XV secolo riferisce invece l'espressione alla vita dell'intera umanità, il cui «mezzo» coinciderebbe con il 1300, data di inizio della composizione, puntellando in questo modo un'interpretazione allegorica del viaggio del pellegrino, in cui vede il percorso dell'intera «humana species». Graziolo Bambaglioli identifica il 1300 con la data di inizio di composizione del poema, mentre Benvenuto da Imola per primo mostra di riconoscere la differenza tra il tempo della finzione letteraria in cui è ambientata la *visio* e quello della stesura dell'opera.
4. Personaggi e soggetti. Il sistema raccoglie i brani in cui compare un medesimo personaggio reale o fittizio (es. Belacqua, Gerione) oppure un medesimo soggetto (es. toponimi ecc.). Es. la ricerca sulla prima occorrenza nel poema del nome di Beatrice (*Inf*., II 70: «I' son Beatrice che ti faccio andare») mostra le differenti prospettive tra gli autori: Iacomo della Lana la identifica in senso esclusivamente allegorico come «scienza teologica»; Iacopo Alighieri la assimila alla «divina Scrittura»; Graziolo Bambaglioli per primo ne riconosce l'identità biografica, limitandosi a dire che fosse «filia condam domini»; infine l'*Ottimo commento*, in posizione isolata rispetto a tutta l'esegesi trecentesca fino a Boccaccio, ne valorizza la doppia funzione storica e allegorica: «alcuna volta pare volere che Beatrice sia quella Biatrice bella che in carne umana elli tanto amòe, e così intendere par volere lo nome alla lettora senza altra allegoria […] alcuna volta pare che lla voglia ponere per beatitudine […] e lo più per la scrittura di teologia».

Il Mediterraneo: un mare di opportunità e sfide

Salvatore Capasso
Università di Napoli "Federico II, Italia
CNR Dipartimento scienze umane e sociali, patrimonio culturale, Italia

ABSTRACT

Il Mediterraneo è un'area caratterizzata da grandi differenze socio-economiche e culturali. Lo shock pandemico prima, e la guerra Russo-Ucraina e l'inflazione poi, hanno accentuato alcune di queste divergenze, ma hanno anche aperto la strada a processi di aggiustamento e di convergenza tra i Paesi delle diverse sponde che fanno di questo mare un'area di grandi opportunità di crescita e sviluppo. Il Mediterraneo è uno degli scenari chiave nel quale si giocheranno le più importanti sfide geopolitiche ed economiche globali come il cambiamento climatico, il calo demografico o la transizione digitale. In particolare, la transizione digitale e il connesso cambiamento tecnologico rappresentano una delle sfide più rilevanti dell'area ma dalla quale possono sorgere grandi opportunitàdi crescita e sviluppo non solo per i Paesi della sponda Sud.

MEDITERRANEO TRA TESTI E CONTESTI

Combining Generative AI and Archaeology to Build Data-Driven Stories

Francesca Buscemi[1], Angelica Lo Duca[2]
[1]National Council of Researches, Institute of Heritage Sciences, Italy - francesca.buscemi@cnr.it
[2]National Council of Researches, Institute of Informatics and Telematics, Italy - angelica.loduca@iit.cnr.it

ABSTRACT[1]

Over the last two years, the spread of Generative AI has opened new opportunities in all the fields. The overall impression is that it could speed up different operations related to content generation, such as narratives and data-driven stories. The article focuses on applying Generative AI to the specific domain of Archaeology. Starting from raw field data and specialized publications, this paper aims to verify the impact of scientific data in the construction of fictional data-driven stories. The proposed system, based on Retrieval Augmented Generation (RAG), combines the archaeological documents and two popular Large Language Models (LLMs): GPT-3.5-turbo and GPT-4.0. Preliminary results show that the implemented system can build quite satisfactory narratives.

KEYWORDS

Generative AI; Data Storytelling; Archaeology.

1. INTRODUCTION

In the last two years, we have witnessed the explosion of Generative AI (GenAI) [6]. This technology enables us to generate new content in the form of texts, images, and audio starting from large quantities of data provided as input. The models used by GenAI, also called Large Language Models (LLMs) [12], are trained offline by technical experts from big companies (such as Google, Microsoft) with data from different sources (often, unfortunately, unknown) and by using a massive amount of computational resources, thus spending prohibitive costs for ordinary people. The trained LLMs can be later used by any user to generate content. Apart from the ethical problems that may arise due to the uncertain origin of data, these models hide enormous potential, which, given their recent development, is not yet known [32, 10].

In this article, we will focus on data storytelling as a specific field of application of GenAI [30] and on its application to Archaeology. Traditionally, data storytelling is used to build non-fictional stories describing insights extracted from data. Data-driven stories are then tailored to a specific type of audience [9].

Otherwise, in this paper, we used data (raw field data and specialized published texts) to build fictional stories based on a real and specific archaeological context. This is to verify the impact of data in constructing AI-assisted data-driven stories addressed to a targeted audience and aimed at the dissemination of archaeological heritage in different environments (educational programs, museum exhibitions, and public outreach activities). We can still define these fictional stories as data-driven stories since they are built starting from data.

According to Joyce [13], contemporary archaeology can be considered "as a discipline engaging in the present in the construction of persuasive stories about imagined pasts." Nevertheless, on one hand archaeologists' selection of details pushes them to show rather than tell [11]; on the other one, the new experiential and sensorial dimension of archaeological heritage fruition, especially in the museums, often mainly focuses on visual narratives more than on textual ones, giving "digital storytelling" a meaning of a story told by the support of digital media, 3D assets, digital design, etc. [7]. On the textual side of storytelling in archaeology, however, crowdsourcing and social media are currently more explored topics [31, 4].

In deciding to switch from technical archaeological data to fictional stories and on the basis of our input to AI, we intended to reflect on two further matters. The first one is the potential of LLMs in endowing the narrative with an emotional tone, able to better engage the public; the second one is the eventual use of biases by the engine, having an impact of an ethical nature. To transform domain-specific scientific documents into fictional data-driven stories, we will use GenAI, and in particular, Retrieval Augmented Generation (RAG) [14]. RAG is a methodology that combines aspects of retrieval-based and generative approaches to natural language processing. RAG refines a generative model with other specific documents or data.

[1] Paragraphs 1, 5 and 6 were written by both authors. Paragraph 2 was written for the first half by Angelica Lo Duca and for the second half by Francesca Buscemi. Paragraph 3 was written by Francesca Buscemi and paragraph 4 by Angelica Lo Duca.

GenAI in Data Storytelling can pose two main ethical concerns: bias and misinformation [22]. Bias in AI involves unjustified preferences or stereotypes in AI systems stemming from altered training data [20]. This can lead to narratives reinforcing stereotypes and compromising the objectivity of data stories. Misinformation may occur when AI generates content that sounds plausible but deviates from reality, potentially spreading misleading information. To tackle these issues, one solution is to review content generated by GenAI tools.

2. RELATED WORK

The literature focusing on combining GenAI and data storytelling is still scarce. Haotian Li et al. defined four figures using GenAI in data storytelling: creator, optimizer, reviewer, and assistant. The *creator* role uses AI to generate data stories from scratch based on the input data and the user's goal [15]. The *optimizer* uses AI to improve the existing data stories by suggesting or applying changes in the content, structure, or style. The *reviewer* uses AI to evaluate and critique the data stories by providing feedback, ratings, or comments. Finally, using AI as an *assistant* involves assisting data workers with various tasks in the data storytelling workflow, such as data collection, cleaning, exploration, modeling, visualization, narration, and presentation. In this paper, we will use GenAI in the creator role.

Lo Duca theorized a possible framework combining GenAI tools and data storytelling to transform data visualization charts into data-driven stories [16]. The same author proposed a RAG-based system to build data-driven stories [17]. Compared with these works, we applied data storytelling to the specific domain of Archaeology and performed some practical experiments.

Sultanum and Srinivasan proposed a framework called DataTales, enabling users to use LLMs to build the textual narrative accompanying data-driven stories [23].

Referring to the generic topic of using LLMs to generate fictional stories, a rising literature exists focusing on different aspects, such as how to build chatbots for storytelling [24], and building a planning model [21]. Zhao et al. used general-purpose LLMs to generate fictional stories and asked roughly 500 participants to rate stories [33]. As a result, they discovered that participants preferred interleaved stories, i.e., stories combining human-generated and LLM-generated content.

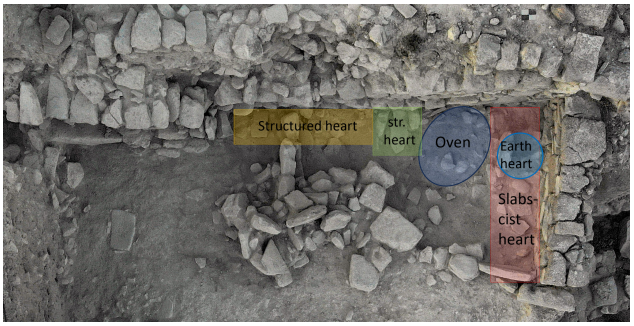
Compared to this study, we use RAG to adapt LLMs to the specific domain of Archaeology. In fact, AI is currently attracting a growing interest [25] in the once technology-phobic environment of archaeology since “The massive digital data in the Humanities is the driving force for the AI applications targeting their study, analysis and interpretation” [19].

Among the sophisticated AI applications in the field of archaeology [29], GenAI is only very recently capturing the attention of archaeologists, being tested in very specialized and technical perspectives such as the use of ChatGPT Language Model for retrieving information in remote sensing and earth observation [1]. Otherwise, the interweaving of GenAI with storytelling in archaeology is much less explored and the two seem still separate. The same word “storytelling” currently appears in connection mainly with serious game plots or IT-assisted narratives conceived for museums and archaeological sites, especially within community-centered projects, because storytelling is considered as a privileged tool for public engagement [4]. Instead, the “thinking” aspect of GenAI calls the archaeologist for caution due to some issues of the engines, highlighted in a very recent article by Peter Cobb [8], such as the plagiarism of training datasets, the lack of factual accuracy, or the human biases of the underlined datasets. Ethical issues made the use of AI in shaping historical narratives controversial [3]. At the same time, the knowledge by ChatGPT of specific archaeological cases (e.g. Late Bronze Age in South Caucaso) [8] has been badly rated, and paved the way to our attempt of data-driven exploitation of GenAI and to all the related challenges that we ourselves observed during the work presented in this paper: targeting the training of these programs on the intellectual output of archaeologists specifically, such as all peer-reviewed publications in the field; combining archaeological evidence and theory; verifying the real creation of new knowledge and opportunities.

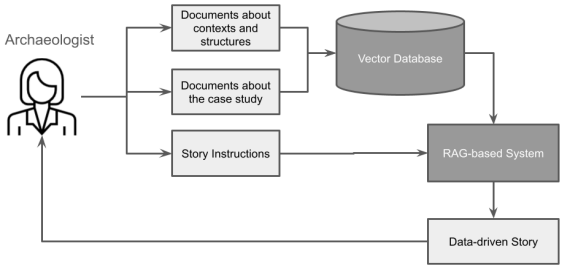
3. ARCHAEOLOGICAL CASE STUDY

Based on our goals, we have chosen a very specialized case study, both with regard to the topographical context and the typology of archaeological evidence. It is located in the site of Phaistos in Crete and focuses on an area excavated in 2022-2023 by the Italian Archaeological Mission at Phaistos directed by the Catania University. This space (Trench 1 to the West of the Room NN) was intended for communal cooking activities, and dated to the Sub-Minoan/Early Iron Age (1050-900 b.C.) (see Fig. 1a). Several fire-installations were discovered leaning against a previous Late Minoan IIIC wall (1200-1050 b.C.) and arranged in a battery: three structured hearts with lithic and clay plates as cooking planes, one great slabs-cist heart as a resumption and enlargement of an earlier earth heart, and one dome clay oven. Unfortunately, the associated pottery does not allow to define a precise chronology for these structures; in any case, they can all be placed between XIth

and IXth century BC. The scanty archaeological levels dating back to the transition between the XIIth and VIII centuries B.C. show the decline of the site in this period and contribute to delineate a very different character of the settlement from the official one of the Minoan period.
As far as the available bibliography is concerned, scientific publications on the matter are definitely rare. In fact, the Iron Age is quite a Dark Age for the Palace of Phaistos, where the impressive Minoan remains have attracted almost exclusively the interest of scholars [5]. The same can be said about the topic of fire-installations of this period at Phaistos and in its territory, where only previous even Neolithic or Bronze Age structures were found [27, 26] except for two Geometric kilns respectively within the Palace and in the very near settlement of Chalara [28].
Within such a scenario, in order to provide LLM a scientific dataset and to test the archaeological accuracy of the GenAI fictional storytelling, we used this kind of texts: very technical texts related to the excavated area and to the findings, such as 2023 excavation reports [18]; general publications concerning Phaistos in the Iron Age (see link to the documents: http://tinyurl.com/2yxwspkt) for the reconstruction of the context and the atmosphere of the stories; articles about fire-installations comparable with our samples, both in chronology and typology, located in other places but Crete, for the structural and usage information. Few articles were related to an earlier period (Bronze Age), given the great typological continuity of these structures.



a)



b)

Figure 1a. Phaistos Palace (Crete). Trench 1, South of Room NN, excavations 2022-2023. Fire-installations, from the Late Minoan IIIC to the Geometric period (XIth-IXth century BC) (F.B.); Figure 1b. The proposed system (Angelica Lo Duca).

4. THE PROPOSED SYSTEM

Figure 1b shows the implemented system used to perform our tests. The system is general and can also be used as a model in other domains. The flow starts with the archaeologist selecting the documents related to the historical context and the focus topic. Both documents can be provided in any format, such as CSV, TXT, PDF, and DOCX.
The *documents about context and structures comparable with the case study* define the historical and archaeological context related to the case study. For example, if we want to generate a story set in the Iron Age, this group of data will include documents describing further comparable information defining a bigger picture and some more general characteristics of that period. This information should help the GenAI model to focus on that period and remove elements from other historical periods.
The *documents about the case study* define the specific case study to analyze. For example, suppose we want to generate a story about kilns from Phaistos. In that case, we should provide the GenAI model system with specific documents about the fire-installations structures in Phaistos, specific data collected during archaeological excavations, and so on.
The archaeologist also defines the *story instructions* to the system. The story instructions are the set of instructions helping the GenAI model to understand what to do with the historical context and focus topic. In this paper, we instructed the GenAI model to generate a fictional tale, but we could instruct it for any purpose.
All the documents are given as input to a *vector database*, which performs indexing and makes them easily and quickly retrievable. In this paper, we have implemented the vector database using Chroma DB[2] and OpenAI GPT-4[3] (GPT4). The vector database and the input defined by the archaeologist are given as input to the RAG-based system, which combines

[2] chroma-core/chroma: "The AI-native open-source embedding database". https://github.com/ chroma-core/chroma.
[3] Official Technical Report on GPT-4: https://doi.org/10.48550/arXiv.2303.08774.

them and produces the *data-driven story*. In this paper, we have implemented the RAG-based system using LangChain[4] and OpenAI GPT-4. The produced story was reviewed by the archaeologist, who can decide to accept it or to modify the story instructions. The process continues until the archaeologist is satisfied with the generated story.

5. EXPERIMENTS AND RESULTS

To use our system we have run seven experiments, each one improving the previous one on the basis of the review by the archaeologist.

Table 1 summarizes the parameters used for each experiment. For all the experiments, except for the last one, we used text-embeddings-ada-002 as the embedding model, i.e., the model used by the vector database. Only in the last experiment we tried GPT-4. In experiments 1-6, we used GPT-3.5-turbo as the LLM for text generation, and in experiment 7, we used GPT-4.

Nr.	Embedding Model	LLM for Text Generation	Documents about contexts and structures comparable with the case study	Documents about the case study	Language of the Output
1	text-embedding-ada-002	GPT-3.5-turbo	Borgna, Corazza, Marchesin (2019)	-	Italian
2	text-embedding-ada-002	GPT-3.5-turbo	Borgna, Corazza, Marchesin (2019)	-	Italian
3	text-embedding-ada-002	GPT-3.5-turbo	Borgna, Corazza, Marchesin (2019)	-	Italian
4	text-embedding-ada-002	GPT-3.5-turbo	All	All	Italian
5	text-embedding-ada-002	GPT-3.5-turbo	All	All	English
6	text-embedding-ada-002	GPT-3.5-turbo	All	All	English
7	GPT-4	GPT-4	All	All	English

Table 1

We started our experiments with a couple of basic trials, by using only one document related to a fire-installation typology in a Bronze Age village at Aquileia, that is to an almost contemporary archaeological context to ours, whose relationship with the ancient Aegean world are attested. No documents on the specific context of Phaistos or Crete were used (Experiments 1-3). We set the output language to Italian, our native language. The output generated by all the experiments is available as additional material associated with this paper (see at the link: http://tinyurl.com/2yxwspkt).

Experiments 1 and 2 have the same story instructions, in order to compare the results: *Generate a 100-word emotional tale in Italian on the topic: FOCOLARI, FORNI E FORNACI TRA NEOLITICO ED ETÀ DEL FERRO.* Both the generated stories, despite insisting on a community-centered significance of fire that could be archaeologically correct for Prehistory/Protohistory, are uncertain about the temporal collocation of the past imagined by the protagonists, fluctuating between the Neolithic and the Iron Age, and rather generic when describing the ancient structures.

In Experiment 3, we changed only the type of tale from emotional to informative. The text is even less satisfying: it tends to take the case study of the input document as the norm, it confuses ovens and kilns with their respective uses and doesn't fit either an outreach purpose.

Experiments 4-7 contain all the specialized aforementioned texts. To increase the engagement degree of the story and to avoid the chronological ambiguity between past and present perceived in the previous experiments, we added in the input that the characters should belong to the period of the case study. We also specified a possible audience for the story. The story instructions in Experiment 4 were: *Generate a 200-word story in Italian addressed to the population of Crete. The characters must be from the period of interest. Topic: FOCOLARI, FORNI E FORNACI.* The generated story did not have the structure of a tale, but rather an educational tone; the narrator speaks in an indefinite period. From an archaeological perspective, only very general information about cooking and fire-installations is provided, again with some confusion between hearth, ovens, kilns, and their respective functions. The archaeological remains described in the story are attributed

[4] Chase, H. "LangChain LLM App Development Framework". https://langchain.com/.

to the entire huge time span of the bibliography used as data input so that the final impression is of a superficial reconstruction of the context; also the site of Phaistos was completely absent and the story lacked a defined setting.
In Experiments 5-6, we tried to better define the story prompts.
In Experiment 5 we used the following story instructions: *Generate a 200-word story about the kiln from Phaistos, told for modern boys living in Crete The characters must be from the Iron Age.* The generated story was improved compared to Experiments 1-4. However, the content of the focus documents was still not present enough.
Thus, in Experiment 6, we decided to be more precise by requesting the model to add details about the Iron Age. We used the following story instructions: *Generate a 200-word story about the kiln from Phaistos. The characters must be from the Iron Age. Add details about the elements of the Iron Age.* In this case, the generated output was the best until now as far as the setting is concerned, with different point of interest: the presence in the storytelling of the great Minoan past of the Phaistos Palace ("towering structures"), the arrangement of the bustling and narrow streets of the Iron Age, the technology advancement in the pottery production due to the introduction of the iron. Even if some inaccuracies are still present (see, for example, the idea of decorative patterns on the kiln), these details are evidently based on the scientific sources we added and provided a credible scenario where the characters act according also to specific emotions and beliefs. In this regard, one point is worth emphasizing: during almost all the XXth century the archaeologists were influenced by a kind of bias towards the post Minoan period in some sites with extraordinary and monumental remains such as Phaistos, up to a point to destroy the later archaeological phases in order to bring to the light the Minoan one.
Experiment 6, on the contrary, reflects a more contemporary and neutral perspective: the kiln/fire represents progress and innovation, the challenges and the crisis of the Iron Age, traditionally framed as a "decadence" period, are occasions for "the resilience and creativity of the people of Phaistos". It could be interesting to verify how much this well-balanced approach to the past was driven by the bibliography we gave the model, in most cases recent, of a technical nature and therefore devoid of ideological perspectives.
As a last chance, we changed the embedding and the text generation models and used GPT-4.
The produced story was comparable to that generated in Experiment 7. It presents a fundamental archaeological error in imaging a kiln in Phaistos aimed at the production of iron object; nevertheless, it expresses, perhaps by a more sophisticated literary canvas and rhetorical devices, all the main issues we observed in Experiment 6: the charm of the Bronze Age civilization, the kiln as a symbol, the technological progress, the solidarity between different generations and therefore the past as a legacy, with a catch final: "it is a story of Phaistos, a city that embraced the future while honoring the past".

6. CONCLUSIONS AND FUTURE WORKS

In this paper we have described the preliminary experiments of applying LLMs and RAG to archaeology to generate narratives from scientific documents tailored to the general public. As a first attempt, we can regard the results as interesting, being a new way to use and reuse data (both archaeological raw data and domain-specific scientific publications) to produce new knowledge, perhaps the main goal of computer applications to Humanities.
At the end of our experiments, the potential of LLMs in transforming these datasets into fictional-driven stories exceeded our expectations.
We needed seven trial tales to reach a quite acceptable pilot result, because we observed that the quality of the outputs strongly depends on the set of instructions given as an input to the system. Thus, by evaluating each tale from time to time, we progressively refined our input, both on the plane of texts given to the Model and on the plane of a clearer definition in our queries of characters, environment, historical period, target audience and literary genre.
We used the observed strong impact of the input on the LLMs storytelling as a way to try to overcome one of the issues highlighted by other scholars about ChatGPT, that is, the human biases of the underlined datasets. Waiting for the the desidered implementation of an explainable AI (XAI), as an instrument for the mitigation of ethical issues [2], we selected dataset texts that were "neutral", directly known to us or recently published and supposed to be free from ideological perspectives. In fact, all our stories resulted inclusive, free from prejudice, exalting values of solidarity between communities and generations, as well as the sense of the past as a legacy, with no stigmatic use of concepts that were present in the older archaeological literature, such as "decadence" or "cultural superiority".
Of course, there is a lot of work to do on the main criticism of GenAI, that is the lack of factual or data accuracy: we obtained a story not yet free of inaccuracies but successful in demonstrating the potential of LLMs to endow a data-driven narrative with an emotional tone, able to engage the general public. Careful and time-consuming work was necessary to avoid the more macroscopic errors in our experiments. On this matter, we think developing a hierarchical organization of the sources for the RAG-based system should improve the system's performance. This is also to "contain" the randomness of the generative process and to increase the impact of the given data.

At the same time, other steps of our proposal can be further developed and reasoned, such as the validation of the described procedure with a greater number of stories, or the assessment of the people's engagement by including a diverse set of human evaluators but the archaeologists/specialists. This could provide a more comprehensive understanding of the story's appeal beyond the domain-specific assessment.

ACKNOWLEDGEMENTS

This work was inspired by the research conducted within the interdepartmental project (DISUM-DMI) "Storage. From data to Web" (Progr. Pia.Ce.Ri, University of Catania, 2021-2024).

REFERENCES

[1] Agapiou, Athos, and Vasiliki Lysandrou. 'Interacting with the Artificial Intelligence (AI) Language Model ChatGPT: A Synopsis of Earth Observation and Remote Sensing in Archaeology'. *Heritage* 6, no. 5 (2023): 4072–4085.

[2] Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, et al. 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI'. *Information Fusion* 58 (2020): 82–115.

[3] Bickler, Simon H. 'Machine Learning Arrives in Archaeology'. *Advances in Archaeological Practice* 9, no. 2 (2021): 186–91.

[4] Bria, Rebecca, and Erick Vasquez Casanova. 'Digital Archaeology and Storytelling as a Toolkit for Community-Engaged Archaeology'. In *Critical Archaeology in the Digital Age*, edited by Kevin Garstki, 2:49–65. Digital Archaeology Series. Cotsen Institute of Archaeology Press at UCLA, 2022.

[5] Buscemi, Francesca. 'Sharing Structured Archaeological Data: Open Source Tools for Artificial Intelligence Applications and Collaborative Frameworks'. *Archeologia e Calcolatori* 34, no. 1 (2023): 145–56.

[6] Cao, Yihan, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 'A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT'. *ArXiv*, 2023. https://doi.org/10.48550/arXiv.2303.04226.

[7] Carlino, Carola. 'Digital Storytelling: Come Rendere Inclusiva La Cultura Attraverso Le Narrazioni Digitali'. *Journal of Inclusive Methodology and Technology in Learning and Teaching* 3, no. 1 suppl. (2023). https://www.inclusiveteaching.it/index.php/inclusiveteaching/article/view/78/72.

[8] Cobb, Peter J. 'Large Language Models and Generative AI, Oh My!' *Advances in Archaeological Practice* 11, no. 3 (2023): 363–369.

[9] Dykes, Brent. *Effective Data Storytelling*. Hoboken: Wiley-Blackwell, 2019.

[10] Frey, Carl Benedikt, and Micheal Osborne. 'Generative AI and the Future of Work: A Reappraisal'. *Brown Journal of World Affairs* 30, no. 1 (2024): 1–12.

[11] Gass, William H. *Fiction and the Figures of Life*. New York: Alfred A. Knopf, 1970.

[12] Gozalo-Brizuela, Roberto, and Eduardo C. Garrido-Merchan. 'ChatGPT Is Not All You Need. A State of the Art Review of Large Generative AI Models'. *ArXiv*, 2023. https://doi.org/10.48550/arXiv.2301.04655.

[13] Joyce, Rosemary. *The Languages of Archaeology*. Hoboken: Wiley-Blackwell, 2002.

[14] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20) Vancouver, 6-12 December 2020. Advances in Neural Information Processing Systems*, edited by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, 9459–9474. 22. Red Hook, NY: Curran Associates Inc., 2020.

[15] Li, Haotian, Yun Wang, Q. Vera Liao, and Qu Huamin. 'Why Is AI Not a Panacea for Data Workers? An Interview Study on Human-Ai Collaboration in Data Storytelling.' *ArXiv*, 2023. https://doi.org/10.48550/arXiv.2304.08366.

[16] Lo Duca, Angelica. 'Towards a Framework for AI-Assisted Data Storytelling'. In *Proceedings of the 19th International Conference on Web Information Systems and Technologies, Rome 15-17 November 2023*, edited by Thomas Bashford-Rogers, Daniel Méneveaux, Mounia Ziat, Mehdi Ammi, Stefan Jänicke, Helen C. Purchase, Kadi Bouatouch, and A. Augusto de Sousa, 512–19, 2024.

[17] Lo Duca, Angelica. 'Using Retrieval Augmented Generation to Build the Context for Data-Driven Stories'. In *Proceedings of the 19th International Conference on Web Information Systems and Technologies, Rome 15-17 November 2023*, edited by Thomas Bashford-Roge, Daniel Meneveaux, Mounia Ziat, Mehdi Ammi, Stefan Jänicke, Helen C. Purchase, Kadi Bouatouch, and A. Augusto Sousa, 690–96, 2024.

[18] Militello, Pietro. 'Scavi Della Missione Archeologica Italiana a Festòs. Le Indagini Delle Campagne 2021-2023'. *Annuario ASAIA* 101, no. 2 (2023): forthcoming.

[19] Pavlidis, George. 'AI Trends in Digital Humanities Research'. *Trends in Computer Science and Information Technology*, 7, no. 2 (2022).

[20] Roselli, Drew, Jeanna Matthews, and Nisha Talagala. 'Managing Bias in AI'. In *Companion Proceedings of the World Wide Web Conference, 13 May, 2019*, edited by Ling Liu and Ryen White, 539–44. New York: Association for Computing Machinery, 2019.

[21] Simon, Nisha, and Christian Muise. 'TattleTale: Storytelling with Planning and Large Language Models'. In *Proceedings of the ICAPS Workshop on Scheduling and Planning Applications, Singapore, 7-12 June 2022*, 2022.

[22] Stahl, Bernd C., and Damian Eke. 'The Ethics of ChatGPT – Exploring the Ethical Issues of an Emerging Technology'. *International Journal of Information Management* 74, no. 102700 (2024). https://doi.org/10.1016/j.ijinfomgt.2023.102700.

[23] Sultanum, Nicole, and Arjun Srinivasan. 'DATATALES: Investigating the Use of Large Language Models for Authoring Data-Driven Articles'. In *Proceedings of the IEEE Visualization and Visual Analytics (VIS), Melbourne 21-27 October, 2023*, edited by Aidan Slingsby, Richard Reeve, and Claire Harris, 231–35. Melbourne, Australia: IEEE Visualization and Visual Analytics (VIS), 2023.

[24] Sun, Yuqian. 'Fictional Worlds, Real Connections: Developing Community Storytelling Social Chatbots through LLMs'. *ArXiv*, 2023. https://doi.org/10.48550/arXiv.2309.11478.

[25] Tenzer, Martina, Giada Pistilli, Ales Brandsen, and Alex Shenfield. 'Debating AI in Archaeology: Applications, Implications, and Ethical Considerations'. *Internet Archaeology* 67 (2024). https://doi.org/10.11141/ia.67.8.

[26] Todaro, Simona. 'Pottery Production at PrePalatial and ProtoPalatial Phaistos: New Data from the Greek- Italian Survey'. In *Αρχαιολογικό Έργο Κρήτης 3: Πρακτικά Της 3ης Συνάντησης Α2, Ρέθυμνο, 5–8 Δεκεμβρίου 2013*, edited by Paulina Karanastasi, Anastasia Tzifkounaki, and Christina Tsigonaki, 495–502. Ephoria delle Antichità di Rethymno, 2015.

[27] Tomasello, Francesco. 'Festòs: Fornace Ad Ovest Del Piazzale I'. *ASAIA* 100, no. II (2023): 9–44.

[28] Tomasello, Francesco. 'La Fornace Del Vano G Nel Quartiere Geometrico Sud-Occidentale Di Festòs'. *Sicilia Antiqua. International Journal of Archaeology* XVI (2019): 193–214.

[29] Traviglia, Arianna. 'Artificial Intelligence Applications to Cultural Heritage'. In Presentation at the 9th Plenary Session of the Steering Committee for Culture, Heritage and Landscape, Council of Europe, Strasbourg, November 4-13, 2020, 2020. https://rm.coe.int/artifical-intelligence-applications-to-cultural-heritage-by-arianna-tr/1680a096b8.

[30] Vidrih, Marko, and Shiva Mayahi. 'Generative AI-Driven Storytelling: A New Era for Marketing'. *ArXiv*, 2023. https://doi.org/10.48550/arXiv.2309.09048.

[31] Vincent, Matthew L. 'Crowdsourced Data for Cultural Heritage'. In *Heritage and Archaeology in the Digital Age. Quantitative Methods in the Humanities and Social Sciences*, edited by Matthew L. Vincent, Víctor Manuel López-Menchero Bendicho, Marinos Ioannides, and Thomas E. Levy, 79–91. Springer, 2017. https://doi.org/10.1007/978-3-319-65370-9_5.

[32] Xu, Minrui, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Shiwen Mao, Zhu Han, et al. 'Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services'. *IEEE Communications Surveys & Tutorials*, 2024, 1127–70. https://doi.org/10.48550/arXiv.2303.16129.

[33] Zhao, Zoie, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P. Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre L.S. Filipowicz. 'More Human than Human: LLM-Generated Narratives Outperform Human-LLM Interleaved Narratives'. In *Proceedings of the 15th Conference on Creativity and Cognition, Virtual Event, 19-21 June 2023*, edited by Léa Paymal and Sarah Fdili Alaoui, 368–70. New York: Association for Computing Machinery, 2023.

Digital preservation e sostenibilità ambientale

Adele Gorini
Università degli Studi di Bologna, Dipartimento di Beni Culturali, Italia – adele.gorini2@unibo.it

ABSTRACT

Il cambiamento climatico si manifesta attraverso eventi estremi, e le nostre azioni individuali devono adattarsi a comportamenti più sostenibili. Sebbene siamo attenti a scelte eco-friendly nella vita quotidiana, l'impatto ambientale derivante dalle attività tecnologiche e dalla conservazione dei dati digitali richiede un'analisi critica. Questo studio esplora l'impatto ambientale delle tecnologie dell'informazione e della comunicazione (ICT) così come della conservazione digitale. Attraverso un'analisi del ciclo di vita degli hardware, dalle materie prime alla produzione, utilizzo e smaltimento, si evidenziano le sfide legate alle emissioni di carbonio. La transizione verso la digitalizzazione ha portato benefici, ma i data centers e la conservazione digitale presentano nuove sfide ambientali. L'utilizzo crescente di servizi cloud ha un forte impatto sulla produzione, gestione e raffreddamento dei data centers, sollevando interrogativi sulla sostenibilità delle pratiche attuali. Si esamina anche il ruolo dell'intelligenza artificiale nell'ottimizzare l'efficienza energetica. L'analisi si conclude sottolineando la necessità di considerare l'impatto ambientale delle attività digitali e di promuovere pratiche sostenibili a livello individuale e aziendale per mitigare l'impatto complessivo sulla crisi climatica.

PAROLE CHIAVE

Sostenibilità digitale; Archiviazione; Cloud; ICT; Digital preservation.

1. INTRODUZIONE

Le dannose conseguenze del cambiamento climatico, manifestatesi attraverso alluvioni, incendi e altre condizioni avverse, sono ormai evidenti in tutte le regioni della Terra. La necessità di comportamenti più responsabili è ormai imperativa e come individui stiamo intensificando i nostri sforzi in diversi settori. Ci impegniamo nella raccolta differenziata, facciamo scelte più sostenibili durante gli acquisiti, optiamo per veicoli elettrici, prediligiamo abbigliamento realizzato con materiali riciclati e cibo biologico. Tuttavia, è altrettanto importante considerare la sostenibilità delle nostre attività digitali. Quanto sono ecocompatibili le nostre scelte in campo tecnologico e digitale? Qual è l'impatto ambientale dei nostri dispositivi e della gestione e conservazione delle nostre memorie digitali? Dal momento che l'interesse per la conservazione non è solo un obiettivo delle comunità archivistiche e culturali, orientate a garantire alle generazioni future l'accesso alle informazioni odierne, ma coinvolge anche i singoli individui, spinti da motivazioni sentimentali, personali, lavorative e organizzative, è cruciale valutare e adottare pratiche sostenibili.

2. METODOLOGIA

La ricerca è stata condotta partendo inizialmente da una revisione della letteratura relativa alle tematiche delle emissioni di carbonio nell'ambito delle tecnologie dell'informazione e della comunicazione (ICT) e conseguentemente nel contesto della conservazione delle memorie digitali. Parallelamente, sono stati esaminati report pubblici provenienti da aziende del settore ICT, con un'attenzione particolare alle iniziative di sostenibilità e agli impatti ambientali derivanti dal ciclo di vita degli hardware e dal consumo dei datacenter.

3. LA CONSERVAZIONE DIGITALE È GREEN?

Dagli ambienti delle grandi società multinazionali, alle piccole medio imprese, dalle istituzioni ai singoli cittadini è evidente come l'intera società sia stata – e sia tutt'ora – protagonista di una irrefrenabile corsa alla digitalizzazione. Senza pressoché alcuna resistenza, il supporto cartaceo è stato - dapprima affiancato e poi - definitivamente sostituito da strumenti tecnologici nella maggior parte dei contesti umani. Per tutti – esperti nel campo e non – è ormai chiaro come la tradizionale disciplina archivistica, legata alla polvere e alla carta, sia giunta al suo termine, almeno per quanto riguarda il contesto della documentazione corrente. [1] Tra le motivazioni alla base della digitalizzazione della società sono state enumerare valide e diversificate ragioni connesse all'accessibilità, alla facilità di condivisione e collaborazione, così come è stato rilevato il suo aspetto ecologico e ambientale, e quindi anche economico. Ma è davvero così? Gli oggetti e i sistemi digitali sono veramente più sostenibili della carta?

[1] Differente rimane il contesto degli archivi storici, almeno secondo chi scrive, per i quali la disciplina tradizionale dovrà a lungo offrire risposte in virtù dei numerosi fondi cartacei che si trovano in stato caotico e ancora non schedati e inventariati.

Da un lato, l'archivistica contemporanea si interroga da tempo sulle nuove modalità di formazione, gestione e conservazione delle memorie digitali e non ha tardato ad esprimere le proprie preoccupazioni in merito, formulando e proponendo le cosiddette *good practice* per sconfiggere l'obsolescenza tecnologica di hardware, software e formati. Dall'altro lato, però, anche se la discussione sulla fragilità dei dispositivi e dei sistemi digitali e sulla *digital preservation* è molto accesa, il dibattito raramente prende in considerazione anche l'aspetto ambientale. Secondo la definizione che ne dà la Digital Preservation Coalition (DPC), la *digital preservation* si configura come «la serie di attività necessarie per garantire l'accesso continuo ai materiali digitali per tutto il tempo necessario»[2] e quindi oltre i limiti dovuti ai guasti dei supporti o dei cambiamenti tecnologici e organizzativi. Sempre secondo la DPC[3], le buone pratiche in campo conservativo si intersecano necessariamente con le considerazioni di sostenibilità ambientale, condividendo un lessico e delle sfide comuni. Entrambe, infatti, mirano ad attribuire agli oggetti e ai sistemi caratteristiche quali quella della durabilità e della persistenza[4], ovvero l'ambizione di scongiurare attacchi da parte di agenti interni e soprattutto esterni che alterino in maniera irreversibile gli oggetti e i sistemi, compromettendone l'originale funzione, stato e fruibilità. Entrambe, poi, condividono la stessa apprensione e preoccupazione per le future generazioni alle quali desiderano consegnare ambienti, sistemi e oggetti preservati e integri. Tuttavia, le pratiche adottate dai singoli così come dalle grandi organizzazioni culturali in ambito di conservazione sembrano dover obbligatoriamente superare i problemi ambientali per garantire l'accesso a una quantità sempre maggiore di dati digitalizzati o nativi digitali. Per meglio spiegare, la conservazione digitale cerca di superare i problemi di fragilità e obsolescenza tecnologica, ad esempio, attraverso la ridondanza degli oggetti e dei sistemi, seguendo la strategia LOCKSS (*Lots of Copies Keep Stuff Safe*): un medesimo contenuto viene salvato in diversi formati e su diversi supporti di memorizzazione e tutte le copie vengono conservate allo stesso modo senza effettuare azioni di selezione e scarto. Questa pratica si riflette anche nel modello OAIS, il quale prevede la creazione e la ricezione di tre pacchetti informativi per il medesimo contenuto digitale, i quali però da un punto di vista funzionale possono esistere come aggregazioni separate e differenti di materiale archivistico[5]. L'insieme di queste pratiche aumenta pertanto la richiesta e il consumo di energia elettrica e quindi l'impronta di carbonio.
Se riflettiamo poi attentamente, l'odierno patrimonio digitale viene creato, consumato e conservato attraverso dispositivi e sistemi creati con le stesse risorse che hanno causato la crisi climatica nel nostro Pianeta. Nello specifico, nell'ambito delle ICT l'impronta energetica può essere causata da tre categorie differenti: dai dispositivi, dai datacenter e dalle reti di comunicazione. Nei paragrafi che seguono si cercherà di analizzare, in particolare, la categoria dei dispositivi e quella dei datacenters.

4. L'IMPATTO DEI DISPOSITIVI

La conservazione digitale deve obbligatoriamente confrontarsi con i supporti di memorizzazione. Tra questi rientrano, per esempio, i nostri smartphone, laptop, hard disk. Per misurare il loro impatto ambientale e la loro impronta di carbonio è necessario prendere in considerazione l'analisi del ciclo di vita dei singoli apparecchi, partendo dalla fase della produzione, passando al trasporto del prodotto nelle mani del consumatore, per arrivare poi alla fase del suo utilizzo da parte dell'utente, il quale lo porterà alla rottura o al deperimento (vd. Fig. 1).

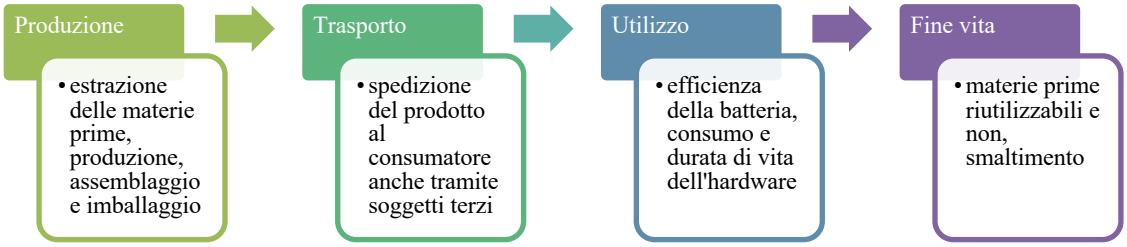


Figura 1. Le quattro fasi del ciclo di vita dei dispositivi

In primo luogo, l'effetto negativo sull'ambiente proviene dalla loro produzione e assemblaggio, per le quali è necessaria l'estrazione e il trasporto di numerose materie prime (tra cui alluminio, cobalto, rame, oro, stagno, litio, vetro, zinco e

[2] Digital Preservation Coalition. "What Is Digital Preservation? - Digital Preservation Coalition." https://www.dpconline.org/digipres/what-is-digipres.
[3] Digital Preservation Coalition. "Environmentally Sustainable Digital Preservation - Digital Preservation Coalition." https://www.dpconline.org/digipres/discover-good-practice/environmentally-sustainable-digital-preservation.
[4] Goldman, Ben. "It's Not Easy Being Green(e): Digital Preservation in the Age of Climate Change," 2019. https://scholarsphere.psu.edu/resources/381e68bf-c199-4786-ae61-671aede4e041.
[5] *Ibidem*.

plastica) [7]. In secondo luogo, l'impatto deriva dal loro trasporto e immissione sul mercato: nel 2021 nell'Unione Europea sono stati movimenti per la vendita ben 13,5 milioni di tonnellate di prodotti elettrici ed elettronici, il cui volume era già di 7,6 milioni di tonnellate nel 2012[6]. In terzo luogo, la loro impronta sull'ambiente proviene ovviamente dal loro utilizzo, proprio e improprio che sia: secondo una ricerca pubblicata nel 2022, le emissioni provenienti da dispositivi informatici e di rete rappresentano circa il 2% delle emissioni totali di carbonio, percentuale che potrebbe addirittura raddoppiare nel prossimo decennio [5]. Si aggiunga poi che l'obsolescenza tecnologica – spesso programmata – rende velocemente i nostri dispositivi inutilizzabili e difettosi, spingendoci all'acquisto di nuovi e al conseguente accumulo di "carcasse" tecnologiche all'interno delle nostre case e istituzioni. Uno studio condotto dalla Commissione Europea[7] ha stimato la presenza di circa 700 milioni di telefoni cellulari e altri rifiuti elettronici depositati nelle famiglie di tutta l'Unione Europea[8]. La consuetudine nel collezionare dispositivi inutilizzati non favorisce logicamente il loro riciclaggio. Vista la scarsità di materie prime e la crescente richiesta globale, il recupero dei materiali assumerebbe un'importanza sempre maggiore nei confronti di un'economia circolare. In questo senso l'Italia, nel 2021 ha prodotto 502660 tonnellate di rifiuti derivanti da apparecchiature elettriche ed elettroniche[9], posizionandosi al terzultimo posto nella classifica dei Paesi Europei che producono questo tipo di rifiuti (vd. Fig. 2). Infine, lo smaltimento improprio di tali apparecchiature può provocare gravi danni ambientali e rischi per la salute umana a causa della presenza di sostanze pericolose.

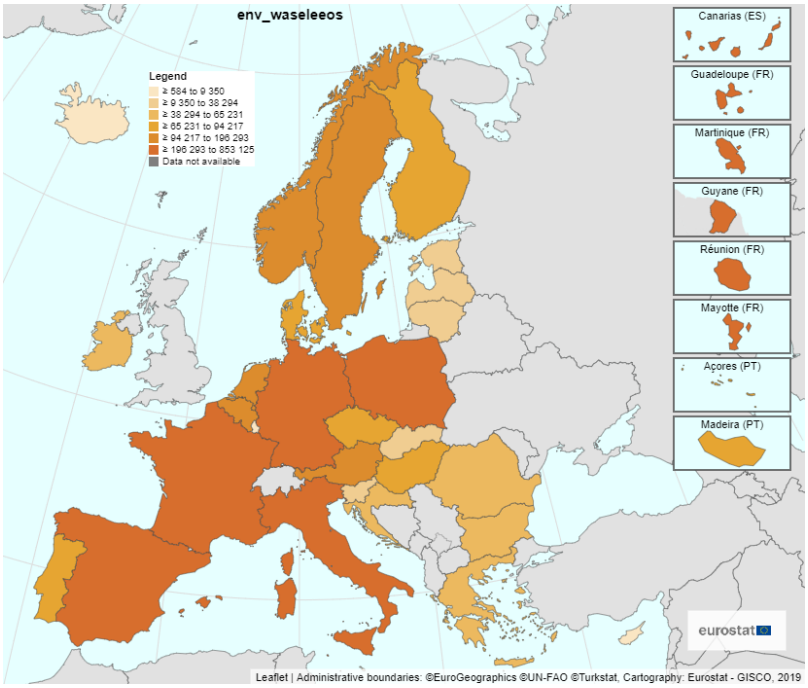


Figura 2. Waste electrical and electronic equipment (WEEE) by waste management operations - open scope, 6 product categories (from 2018 onwards)

5. IL CLOUD: LA NUVOLA 'CARNIVORA'

Sempre più utenti e organizzazioni si affidano ai servizi cloud per conservare le proprie memorie in modo affidabile e duraturo. Il termine stesso "cloud", evocativo di qualcosa di immateriale e leggero, suggerisce un'idea di distacco e trascendenza. Tuttavia, come ben sappiamo, si tratta di una tecnologia estremamente materiale e massiccia, basata

[6] Eurostat. "Waste statistics - electrical and electronic equipment". October 2023. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Waste_statistics_-_electrical_and_electronic_equipment - Electrical_and_electronic_equipment_.28EEE.29_put_on_the_market_and_WEEE_processed_in_the_EU.
[7] European Commission. "Commission Recommendation (EU) 2023/2585 of 6 October 2023 on improving the rate of return of used and waste mobile phones, tablets and laptops". 2023.https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AL_202302585.
[8] A parere di chi scrive, sarebbe interessate, in questo senso, ampliare la ricerca anche nei confronti dei rifiuti elettronici presenti nelle istituzioni pubbliche.
[9] European Commission. "Waste from Electrical and Electronic Equipment (WEEE)". 2021. https://environment.ec.europa.eu/topics/waste-and-recycling/waste-electrical-and-electronic-equipment-weee_en.

sull'utilizzo di data centers. In effetti, lo studioso Steven Gonzales Monserrate, per descrivere questa tecnologia, ha coniato il termine "carbonivora"[10] proprio in riferimento all'impatto ambientale che deriva dal suo utilizzo.
La transizione dalla conservazione tradizionale dei documenti cartacei a quella digitale offre indubbi vantaggi, come la riduzione dell'uso di carta, il risparmio di spazio interno e la possibilità di accesso da remoto. Tuttavia, è essenziale considerare che alcune condizioni fondamentali rimangono immutate. Una di queste è il luogo in cui vengono costruiti i depositi, che deve essere attentamente scelto indipendentemente dal tipo di contenuti che ospita (cartacei e digitali). La struttura deve essere progettata considerando l'ambiente circostante, l'esposizione al sole e la vicinanza a corsi d'acqua – si pensi ai danni provocati dall'alluvione che ha colpito il territorio forlivese e l'archivio di deposito comunale[11] – o aree sismiche. In secondo luogo si deve tenere in considerazione anche l'impatto ambientale derivante dalla costruzione stessa del deposito digitale o cartaceo, la quale può essere responsabile di una grave impronta di carbonio, anche detta 'impronta incorporata'.
Un'analisi dettagliata deve poi essere rivolta nei confronti dei consumi energetici dei data centers. Come nei depositi cartacei, dove è necessario mantenere certe temperature e livelli di umidità per conservare i materiali analogici, anche nei data centers è essenziale mantenere temperature fresche e costanti per evitare il surriscaldamento delle componenti elettroniche [6]. Per il raffreddamento, le aziende possono scegliere tra due tipi di sistemi: ad aria e ad acqua. Tuttavia, il raffreddamento ad acqua risulta preferibile in quanto più efficiente ed economico, con un risparmio energetico del 10% rispetto al raffreddamento ad aria e quindi significativamente meno impattante a livello carbonico. Le aziende ITC affermano di combattere l'impatto ambientale utilizzando energia proveniente da fonti rinnovabili come eolica e solare. Tuttavia, l'accessibilità a queste energie può essere limitata, e i data centers potrebbero dover ricorrere a fonti energetiche non rinnovabili in caso di carenza.
Infine, l'impatto ambientale deriva dalla richiesta di accesso – seppur da remoto – ai documenti ivi conservati e al trasferimento e immissione di nuovi dati da conservare. Anche se attualmente i dati in nostro possesso non sembrano essere allarmanti in questo senso, è comunque bene sottolineare che l'impronta di carbonio derivante dalla conservazione di documenti sul cloud aumenterà nei prossimi anni a fronte del crescente volume di dati digitali prodotti e della mancanza di pratiche di scarto in ambiente digitale[12].

6. SOLUZIONI E IMPEGNO

L'attenzione dell'Unione Europea è sempre più concentrata nei confronti delle questioni ambientali e nel corso degli anni la Commissione si è fatta promotrice di direttive e regolamentazioni per ridurre l'impronta di carbonio sotto diversi punti di vista. Sicuramente da citare è la *WEEE Directive*, ovvero la direttiva sui rifiuti di apparecchiature elettriche ed elettroniche, che da almeno un decennio promuove la raccolta, il recupero e il riciclaggio dei rifiuti elettronici, incoraggiando una gestione più sostenibile dei materiali non solo a livello industriale ma anche individuale[13]. Più recentemente l'Unione, anche a seguito dell'Accordo di Parigi del 2015 [3], ha intrapreso una nuova strategia per rendere l'Europa più efficiente sotto il profilo delle risorse grazie al *Green Deal*. Esso rappresenta una sfida ambiziosa, volta a rendere l'Europa il primo continente climaticamente neutro entro il 2050. Presentato nel 2019, il *Green Deal* richiede la collaborazione di tutti i settori della società (governi, imprese, istituzioni e cittadini) per la promozione di energie rinnovabili e dell'efficienza energetica, così come la transizione verso un'economia circolare che favorisca il riciclo, evitando gli sprechi[14].
Queste iniziative europee sono un chiaro segnale di attenzione nei confronti dell'ambiente al quale anche le più importanti aziende del settore non hanno tardato ad aderire. Al primo posto si posiziona probabilmente Google, il quale è tra i principali detentori di datacenters e servizi cloud insieme con Amazon. Il colosso statunitense si professa come il primo ad aver

[10] Gonzalez Monserrate, Steven. "The Staggering Ecological Impacts of Computation and the Cloud." Scientific American, March 1, 2022. https://www.scientificamerican.com/article/the-staggering-ecological-impacts-of-computation-and-the-cloud/.
[11] Rai news. "L'archivio Comunale Di Forlì Devastato Dall'alluvione." Rai, June 21, 2023. https://www.rainews.it/tgr/emiliaromagna/articoli/2023/06/larchivio-comunale-di-forli-devastato-dallalluvione-9af4d0fa-bcc0-4b69-be77-cc5ebeb4f854.html.
[12] Addis, Matthew. "Does net zero emissions from energy usage in the cloud mean carbon free digital preservation is on the horizon?". 2023. https://www.dpconline.org/blog/blog-matthew-addis-enviornmental-23.
[13] EUR Lex. "Consolidated text: Directive 2012/19/EU of the European Parliament and of the Council of 4 July 2012 on waste electrical and electronic equipment (WEEE) (recast) (Text with EEA relevance)". https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A%3A02012L0019-20180704.
[14] European Commission. "Il Green Deal europeo". 2019. https://eur-lex.europa.eu/resource.html?uri=cellar:b828d165-1c22-11ea-8c1f-01aa75ed71a1.0006.02/DOC_1&format=PDF.

accolto la sfida green, diventando nel 2007 la prima grande azienda *carbon neutral*, e promette ora di raggiungere «un'energia priva di emissioni di carbonio ovunque, in ogni momento, entro il 2030»[15].
Un altro esempio tangibile è rappresentato da Apple, che ha mostrato un impegno concreto nella riduzione delle emissioni di carbonio. Attraverso i suoi *Environmental Progress Report* [2], Apple ha documentato una significativa riduzione della sua impronta di carbonio, diminuendo del 19,12% rispetto al 2019 e del 47,14% rispetto al 2015 [1].
A dimostrare una seria presa di coscienza e un impegno verso pratiche più sostenibili e rispettose dell'ambiente, non sono state solo le grandi aziende ma anche alcune istituzioni. Tra queste si segnala, per esempio, The National Archives of the Netherlands, la quale, accogliendo la sfida proposta dalla Digital Preservation Coalition, ha incluso una sezione sulla sostenibilità ambientale nella propria politica di conservazione digitale, mirando alla creazione di un archivio digitale neutro dal punto di vista climatico[16].
Recentemente è stato anche proposto l'impiego dell'intelligenza artificiale (AI), intesa come l'insieme di tecniche matematiche e informatiche mirate all'analisi dei dati per aiutare a comprendere e navigare fenomeni del mondo reale, offrendo previsioni migliori e suggerendo strategie per ottimizzare i risultati. L'AI in questo senso può essere un utile strumento per aiutare a gestire i problemi derivanti dal cambiamento climatico, grazie alla sua capacità di raccogliere, analizzare e interpretare ampi set di dati. Secondo uno studio condotto dall'UC Berkeley e Google nel 2022, i datacenter rientrano in una della quattro aree in cui l'AI può aiutare a ridurre l'utilizzo di energia e di emissioni, in linea con l'obiettivo di eliminarle definitivamente entro il 2050 [4]. Anche the Royal Society si è espressa in questo senso, sottolineando come l'intelligenza artificiale potrebbe rivelarsi lo strumento più facile per giungere ad un greening computing dal momento che aiuterebbe ad ottimizzare non solo il riscaldamento e raffreddamento dei datacenters ma anche la progettazione di hardware e l'uso delle strutture informatiche per evitare server inattivi [9]. In sostanza, l'AI sarebbe applicabile sia nel campo della mitigazione - poiché in grado di ridurre le emissioni di gas serra di un'organizzazione del 5% al 10% rispetto alla sua impronta di carbonio attuale – sia nell'ambito della capacità di adattamento e di resilienza, aumentando la protezione dai rischi di effetti sul lungo periodo o di eventi estremi. L'ostacolo principale nei confronti dell'adozione di questo strumento deriva in primo luogo dalla mancanza di competenze e in secondo luogo alla mancanza di fiducia nei suoi confronti. Pertanto, solo attraverso un sostegno maggiore - sia a livello economico sia a livello di risorse umane impiegate – l'AI potrà produrre soluzioni adeguate e aiutare le organizzazioni e le aziende del settore a migliorare i loro livelli di produzione [8].

7. CONCLUSIONE

In conclusione, le sfide ambientali derivanti dal cambiamento climatico richiedono una risposta concreta e globale da parte di tutti i settori della società, compreso quello della conservazione digitale. La nostra dipendenza dalla tecnologia ha un impatto significativo sull'ambiente, con emissioni di carbonio provenienti dalla produzione, utilizzo e smaltimento dei dispositivi digitali, nonché dall'energia consumata dai data centers. Tuttavia, esistono soluzioni e strategie che possono contribuire a mitigare questo impatto, aiutandoci ad adottare pratiche più sostenibili.
Le iniziative europee come la *WEEE Directive* e il *Green Deal* evidenziano un impegno concreto verso la sostenibilità ambientale, mentre alcune aziende e istituzioni stanno adottando misure per ridurre le proprie emissioni di carbonio e promuovere la conservazione digitale sostenibile. L'introduzione dell'intelligenza artificiale come strumento per ottimizzare l'efficienza energetica nei data centers e ridurre le emissioni di carbonio rappresenta un'ulteriore opportunità per affrontare questa sfida. Tuttavia il ricorso a questa tecnologia non deve distogliere tuttavia i singoli individui dall'impegno proattivo nella lotta al cambiamento climatico.
La sostenibilità ambientale nel settore dei dati e della tecnologia, infatti, non deve essere solo una priorità delle multinazionali e delle istituzioni culturali delegate alla custodia e alla conservazione. Per una riuscita ottimale della sfida verde, è necessario il contributo di ognuno di noi: tramite scelte e acquisti consapevoli e la messa in atto di buone pratiche nell'utilizzo degli strumenti informatici possiamo giocare un ruolo determinate nell'abbattimento delle fonti non green e nella ridefinizione di un'economia digitale attenta all'ambiente.

[15] Google Cloud Blog. "Google Cloud aims for carbon-free energy for its data centers". 2020. https://cloud.google.com/blog/topics/inside-google-cloud/announcing-round-the-clock-clean-energy-for-cloud.
[16] Van Hoek, Sophia**,** "Walking a tightrope across the gap of digital preservation and environmental sustainability: The National Archives of the Netherlands and the challenge of achieving a climate-neutral digital archive". 2023. https://www.ahk.nl/media/ahk/docs/eindwerkprijs/2023/Walking_a_tightrope.pdf.

BIBLIOGRAFIA

[1] Apple. ‘Environmental Progress Report’, 2020. https://www.apple.com/euro/environment/pdf/a/generic/Apple_Environmental_Progress_Report_2020.pdf.

[2] Apple. ‘Environmental Progress Report’, 2023. https://www.apple.com/environment/pdf/Apple_Environmental_Progress_Report_2023.pdf.

[3] Cook, Gart. ‘Clicking Clean: A Guide to Building the Green Internet. May 2015.’ Greenpeace, 2015. https://www.greenpeace.org/static/planet4-international-stateless/2015/05/153e0823-2015clickingclean.pdf.

[4] Dannouni, Amane, Stefan A. Deutscher, Ghita Dezzaz, Adam Elman, Antonia Gawel, Marsden Hanna, Andrew Hyland, et al. ‘Accelerating Climate Action with AI’, 2023. https://www.gstatic.com/gumdrop/sustainability/accelerating-climate-action-ai.pdf.

[5] Freitag, Charlotte, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon S. Blair, and Adrian Friday. ‘The Real Climate and Transformative Impact of ICT: A Critique of Estimates, Trends, and Regulations’ 2, no. 9 (2021). https://doi.org/10.1016/j.patter.2021.100340.

[6] Glanz, James. ‘Power, Pollution and the Internet’. *The New York Times*, 23 September 2012. https://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html.

[7] Gupta, Udit, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. ‘Chasing Carbon: The Elusive Environmental Footprint of Computing’. *ArXiv*, 2020. https://doi.org/10.48550/arXiv.2011.02839.

[8] Maher, Hamdi, Hubertus Meinecke, Damien Gromier, Mateo Garcia-Novelli, and Ruth Fortmann. ‘How AI Can Be a Powerful Tool in the Fight Against Climate Change’, 2022. https://web-assets.bcg.com/ff/d7/90b70d9f405fa2b67c8498ed39f3/ai-for-the-planet-bcg-report-july-2022.pdf.

[9] The Royal Society. ‘Digital Technologies and the Planet. Note of Discussions at a Royal Society Workshop, 26 November 2019’, 2019. https://royalsociety.org/-/media/policy/projects/digital-technology-and-the-planet/digital-technology-and-the-planet-summary-notes.pdf.

Il progetto digitale su Elisa Chimenti (LAI-ALEEF): la problematicità di un profilo mediterraneo tra reti e frontiere

Ada Desideri[1], Bianca Vallarano[2]
[1] Sorbonne Université, Francia - ada.desideri@sorbonne-universite.fr
[2] Università di Napoli L'Orientale/Université de Lille, Italia/Francia - b.vallarano@unior.it

ABSTRACT
Il paper presenta i lavori in corso sull'opera e sull'archivio di Elisa Chimenti (Napoli 1883-Tangeri 1969), per proporre delle riflessioni metodologiche sulle problematicità dei progetti digitali che implichino una rete trans-mediterranea. In primo luogo, viene presentata l'autrice e il suo corpus alla luce della sua pluri-appartenenza geografica e linguistica, che la inserisce a pieno titolo in una complessa rete mediterranea. In seguito, viene presentato il progetto Chimenti, un progetto di digitalizzazione del corpus che tenta di ovviare grazie al digitale agli impedimenti ed ai limiti geografici, metodologici e scientifici che un corpus tale presenta, dal plurilinguismo alla standardizzazione della codifica, dall'interdisciplinarietà alla transnazionalità. Infine, sono evidenziate le problematicità che rimangono aperte, tanto da un punto di vista teorico quanto pratico, riguardo agli strumenti e alle metodologie del digitale.

PAROLE CHIAVE
Elisa Chimenti; plurilinguismo; spazio mediterraneo; archivi digitali; edizioni digitali.

1. INTRODUZIONE
Parlare di Mediterraneo non vuol dire solo parlare di reti, ma soprattutto di frontiere. Spazio transnazionale, fin dall'antichità luogo di incontro e scontro dei popoli che lo hanno abitato, il bacino Mediterraneo è rappresentato a seconda delle epoche e dei punti di vista come culla e come campo di battaglia, come ponte e come muro [18]. Ogni rappresentazione dello spazio mediterraneo dipende da ed è costruita su presupposti storico-culturali differenti, iscrivendosi in una specifica genealogia storico-culturale [14]. Attraversare il Mediterraneo, concretamente e metaforicamente, è un'operazione ancora oggi non scontata, ed asimmetrica [9]. Fare storia del Mediterraneo, fare letteratura del Mediterraneo, vuol dire ancora oggi trovarsi davanti continue frontiere [4]. Il caso di Elisa Chimenti si inserisce in tale panorama scientifico-disciplinare, e può essere considerato esemplificativo di una serie di tensioni che vi sono connaturate, tanto da un punto di vista geo-storico quanto metodologico. Cosa significa provare a fare rete nello spazio mediterraneo? Quali sono le difficoltà e le tensioni che si incontrano? Concretamente, come tali problematiche emergono e si ritrovano nei progetti digitali sul corpus? La riflessione si inserisce dunque nel secondo asse della *call for papers*, affrontando nel concreto temi e questioni inerenti alle edizioni scientifiche digitali nonché la rappresentazione e l'organizzazione delle tecniche e criticità nel campo della digitalizzazione, dell'organizzazione e dell'archiviazione del patrimonio, nel contesto della rete dei paesi gravitanti sul Mediterraneo.

2. ELISA CHIMENTI: UNA AUTRICE MEDITERRANEA
Elisa Chimenti è un chiaro esempio di un'autrice che a causa della sua pluri-appartenenza, nonché nel suo genere, non è entrata in nessun canone nazionale e ha avuto difficoltà, in vita come *post mortem*, a vedere riconosciuta la propria autorialità [6]. Di origine napoletana ma emigrata con la famiglia all'età di pochi mesi nel Maghreb, Chimenti ha vissuto e operato nel Marocco coloniale e postcoloniale, e specificamente nella città di Tangeri, zona internazionale fino all'indipendenza del paese ottenuta nel 1956. Italiana, Chimenti perde la cittadinanza ed assume quella tedesca in seguito al matrimonio, avvenuto a Tangeri nel 1912. Apolide dopo il divorzio, ha intrattenuto per tutta la vita relazioni problematiche con le istituzioni dei tre paesi. Autrice poliglotta, si è formata nello spazio internazionale e mediterraneo della regione di Tangeri, fin dall'antichità crocevia di lingue e popoli diversi [19, 2]. La sua opera è profondamente *métissée*, tanto da un punto di vista linguistico quanto di tradizioni ed immaginari [17]: nei suoi testi, fortemente ancorati nell'oralità, le diverse lingue del bacino mediterraneo si intersecano, dando voce ai popoli che lo hanno attraversato [28]. Così, sulla base francese risuona l'arabo e specificamente la *darija* marocchina, che si mescola ai dialetti italiani del Sud, all'ebraico sefardita, allo spagnolo vivo nello stretto di Gibilterra, all'amazigh della catena montuosa del Rif. Il risultato è

un ibridismo [5] che sfugge ad ogni definizione e categorizzazione univoca, presentandosi come un mosaico di storie letterarie, orali, culturali e religiose tanto distinte quanto strettamente interrelate [7].
Nonostante Chimenti abbia scritto moltissimo e collaborato con diversi giornali marocchini ed internazionali, ha pubblicato in vita solamente cinque opere [10][1]. Migliaia di pagine rimangono inedite, conservate presso la *Fondation Méditerranéenne Elisa Chimenti*[2]. L'associazione, fondata nel 2010 con sede presso il *Palais des Institutions Italiennes* di Tangeri, è nata grazie all'impulso di intellettuali ed amici vicini all'autrice, con lo scopo di promuoverne l'opera. Purtroppo, nella mancanza di un ente o di una istituzione pubblica che garantisse la durabilità del progetto, le attività della fondazione sono da alcuni anni cessate, e la sistematizzazione dell'archivio non è stata terminata. La sede della Fondazione è oggi chiusa, e l'archivio non accessibile. Senza dubbio, il fatto che l'archivio interpellasse interlocutori nazionali diversi ha contribuito alle difficoltà nella sua sistematizzazione e durabilità: cittadina tedesca e italiana, Chimenti ha lavorato per tutta la sua vita in Marocco con gli enti italiani esteri (la scuola italiana, da lei fondata nel 1914), che, dopo la fine del periodo coloniale, sono stati definitivamente chiusi [27]. La condizione frammentata degli archivi del periodo coloniale è tutt'oggi una questione viva, e problematica [11].
Tanto la vita quanto l'opera e i lasciti di Elisa Chimenti si articolano dunque in una complessa rete mediterranea, che tenta di superare le frontiere nazionali per istituire un dialogo trasversale. È questa una delle cause che ha determinato l'esclusione dell'autrice da un canone che ancora oggi si identifica troppo spesso con un identitarismo eurocentrico nazionale e monolingue, ed è reticente ad accogliere identità plurali e composite [3, 26]. Il riconoscimento dell'autorialità andando di pari passo con l'istituzionalizzazione degli archivi – che non sono semplice luogo fisico ma luogo simbolico, che istituzionalizza una legge sull'altra [15] – il risultato è un circolo vizioso che permette ad un sistema chiuso di perpetuarsi, *ne varietur*. Il lavoro archivistico di preservazione, studio, organizzazione dei documenti è il primo passo per smuovere il canone e diffondere nuovi paradigmi di conoscenza. In questo senso, la riflessione sugli archivi non è una riflessione sul passato, ma sul presente – in quanto luogo di educazione e mobilitazione [13] – e sul futuro.

3. IL CORPUS CHIMENTI: I PROGETTI DIGITALI

Di fronte alle sfide poste da un archivio non sistematizzato, un corpus variegato ed una rete vasta e complessa che interpella interlocutori diversi, il digitale dovrebbe aprire maggiori possibilità rispetto ai metodi tradizionali. È per questo che i progetti di valorizzazione di tale corpus hanno intersecato fin da subito lo studio accademico e la preparazione delle edizioni a stampa degli inediti con progetti e approcci digitali. In un primo momento, questo aspetto è stato sviluppato nell'ambito del Laboratorio Associato Internazionale (LAI) "La scrittura dell'esilio al femminile. Il dialogo tra le lingue, le culture e le idee nello spazio europeo e mediterraneo XIX-XXI secolo" (2018-2022)[3]. In questo contesto, nato dalla collaborazione tra università diverse in Francia, Italia e Spagna, il lavoro informatico si è focalizzato sulla concezione e la creazione di edizioni critiche digitali, con lo scopo di diffondere i testi in ambito di ricerca, di salvaguardarli, ma anche e soprattutto di riuscire a rappresentare la loro complessità grazie alle visualizzazioni informatiche [8][4].
Un primo esperimento di edizione digitale [12] è stato effettuato su due testi molto diversi per forma e contenuto: un romanzo di fantascienza in bella copia (*Une étrange aventure*) e la bozza di un racconto ambientato durante la guerra del Rif (*Mennouch la Rifaine*, della raccolta *Contes Berbères*). In questi due progetti, l'annotazione si basa su un documento – concepito a partire da un'analisi generale del corpus e dei temi affrontati dall'autrice[5] – che aveva l'obiettivo di rappresentare un massimo di fenomeni ritenuti interessanti, senza distinzioni tra informazioni sulla struttura del testo ed elementi di analisi. Queste edizioni permettono di leggere due versioni del testo (semi-diplomatica e critica); di visualizzare parole e passaggi relativi a temi ricorrenti nel corpus; di creare delle liste a partire da queste parole; di far apparire la lingua utilizzata per le parole straniere; e di leggere le note legate all'apparato critico.

[1] Si tratta di: *Èves marocaines* (Tanger, 1935), *Chants de femmes arabes* (Paris, 1942), *Au cœur du harem* (Paris 1958), *Légendes marocaines* (Paris 1959), *Le sortilège et autres contes sephardites* (Tanger, 1964).
[2] Cfr. il sito internet della Fondazione: https://www.elisachimenti.org/accueildef.html.
[3] Progetto internazionale coordinato da Camilla Cederna (Università di Lille, lab. CECILLE) e Silvia Tatti (Università Sapienza di Roma), in collaborazione con l'Università di Pisa e l'Università di Siviglia. I due casi di studio focus del progetto erano Elisa Chimenti e Cristina Trivulzio di Belgiojoso (https://cecille.univ-lille.fr/vie-scientifique).
[4] Le edizioni sono depositate su GitLab, e possono essere condivise individualmente su richiesta a scopo di ricerca. Per motivi di diritto d'autore, al momento non possono essere rese pubbliche. Per questo motivo, non abbiamo potuto includere in questo articolo le immagini delle interfacce delle due edizioni.
[5] Il documento è stato creato da Camilla Cederna, Nathalie Gasiglia e dalla studentessa Camélia Ait-Mechedal.

Un secondo esperimento di edizione digitale è stato preparato su un racconto che presenta delle varianti genetiche (*Histoire du prince Moustapha et de la princesse Djamila*[6]), utilizzando il software di visualizzazione EVT. L'annotazione si è focalizzata – oltre che sull'arborescenza e su alcune annotazioni di base per esplorare il testo (persone, luoghi, lingue straniere) – sulla costruzione di un apparato critico, con l'obiettivo di rendere conto degli aspetti genetici. Si sono quindi combinati i due moduli TEI *basic apparatus* (per note, aggiunte, cancellature) e *critical apparatus* (per testimoni, lemmi, readings), entrambi codificati con il *parallel segmentation method*. Grazie a EVT, il testo è consultabile in diverse modalità di edizione (testo semplice, edizione fotografica, collazione). È possibile evidenziare persone, luoghi e parole in lingua straniera, creare delle liste a partire da questi elementi, e visualizzare i fenomeni legati allo sviluppo genetico del testo.
Parallelamente ai lavori di edizioni digitali, un nuovo progetto è iniziato nel 2023, finanziato dalla *Maison Européenne des Sciences de l'Homme et de la Société* (MESHS)[7], dal titolo *Archives de l'écriture de l'exil au féminin* (ALEEF)[8]. Il progetto, che coinvolge il corpus di Elisa Chimenti, si pone come obiettivo di proseguire la ricerca iniziata in seno al LAI e portare avanti il lavoro sul digitale. In questa ottica, è nata la collaborazione con il laboratorio di ricerca Thalim[9] (CNRS, ENS, Sorbonne Nouvelle) per costruire una biblioteca digitale dell'autrice, gli *Archives Chimenti*, online sulla piattaforma EMAN[10]. Un altro obiettivo del progetto è stato quello di ampliare la rete di collaborazioni alla sponda meridionale del Mediterraneo: alle università in Europa si sono aggiunte università e enti in Marocco[11], con i quali la collaborazione è in corso sia per lo studio linguistico e letterario del corpus, sia per la riflessione sullo stato dell'archivio.

4. RIFLESSIONI SUL CAMPO OVVERO LE PROBLEMATICHE INCONTRATE NEI PROGETTI DIGITALI

Rispetto agli obiettivi e ai risultati attesi da tali progetti digitali, sono presto emerse le problematiche dovute al lavoro in un contesto ibrido, sia sul piano tecnico/materiale sia sul piano dei processi di formalizzazione dei testi.
Dal punto di vista tecnico, i problemi più evidenti sono legati all'OCR. In tutti gli esperimenti di edizione presentati, è stato fatto un primo tentativo di trascrizione usando dei software di OCR, prima di effettuare la trascrizione manuale. I risultati dei software usati (*Tesseract* e *Abbyy Fine Reader*) non erano soddisfacenti, anche per i testi in bella copia e ben conservati. I motivi sono due, ed entrambi dovuti alla complessità di una rete vasta, non organizzata e plurilingue. Da una parte, la bassa qualità delle immagini, diretta conseguenza della difficoltà di comunicazione tra paesi e istituzioni diverse – il che comporta la non sistematizzazione dell'archivio – e della conseguente insufficienza delle risorse. Le fotografie disponibili dei testi sono state scattate con i telefoni personali di due ricercatrici, che hanno potuto accedere al fondo una prima volta tra il 2018 e il 2019 ed una seconda volta nel 2022. Una scannerizzazione professionale era prevista, ma non ha potuto essere completata vista l'inaccessibilità del fondo. D'altra parte, seppure avessimo a disposizione immagini di buona qualità, ulteriore problema è l'ibridismo linguistico dei testi: i software di OCR sono ottimizzati per ricevere una sola lingua in input, e generano moltissimi errori in contesti plurilingui. A questo si aggiunge, ad un livello ulteriore, la problematicità delle lingue trattate. Tali software sono addestrati solo sulle lingue scritte e maggiormente sulle lingue occidentali politicamente maggioritarie[12]. Infine, neanche un OCR addestrato sulla *darija* del Marocco permetterebbe di risolvere questo problema, poiché i termini arabi sono traslitterati da Chimenti in alfabeto latino in maniera eterogenea. Abbiamo quindi la certezza che le parole in lingua straniera e soprattutto quelle in arabo e *darija* marocchina – che costituiscono una percentuale importante del lessico di Chimenti – non sarebbero trascritte correttamente.
Per quanto riguarda la rappresentazione formale dei testi, la pluralità che caratterizza il corpus ci pone di fronte a difficoltà metodologiche ed epistemologiche, legate allo specifico progetto di edizioni digitali. Per quanto riguarda le prime tipologie

[6] Il racconto è presente in due raccolte dattiloscritte di Chimenti, conservate entrambe nel suo archivio: *Contes d'autrefois* (con correzioni manoscritte) e L*a veillée du harem* (in bella copia).
[7] CNRS, UAR 3185: https://www.meshs.fr/page/Archives_Ecritures_Exil_Feminin....---titre-------asc.
[8] Archivi della scrittura dell'esilio al femminile, il progetto è coordinato da Camilla Cederna: https://lai-ecriture-exil-au-feminin.univ-lille.fr/accueil.
[9] *Théorie et histoire des arts et des littérature de la modernité XIXe-XXie siècle*, CNRS, UMR 7172: https://www.thalim.cnrs.fr/.
[10] EMAN (*Éditions de manuscrits et d'archives numériques*) è una piattaforma di pubblicazione digitale per la diffusione di manoscritti e di fondi d'archivio moderni. Si basa sul software Omeka, ma prevede numerose estensioni: https://eman-archives.org/EMAN/.
[11] L'Università Abdelkmalek Essaâdi di Tetouan, il Centro culturale Beit Nedjma, gli Archives du Maroc.
[12] Il problema dei *bias* occidentali negli strumenti di DH è ben più ampio [29, 24]. D'altronde, i sistemi riflettono i *bias* delle società che li hanno creati. L'esperienza dimostra come gli strumenti di DH (dagli OCR ai database agli strumenti di *text analysis*) lavorano meglio con le lingue occidentali (e in particolare con l'inglese) che con le lingue definite "minori" (che spesso in realtà minori non sono, ma piuttosto marginali politicamente, cfr. [16]). Sono molte oggi le ricerche che si muovono per sopperire a queste mancanze. Cfr. non esaustivamente [22, 25, 1], negli atti della *7th IEEE Congress on Information Science and Technology (CiSt)*, Agadir - Essaouira, Morocco, 2023.

di edizioni, la volontà di rappresentare informazioni su aspetti diversi – strutturali ed analitici – ha fatto emergere una serie di problematiche. Oltre al tempo richiesto per trattare un grande numero di fenomeni, le annotazioni sono risultate talvolta non pertinenti e talvolta insufficienti rispetto ai nostri obiettivi. Da una parte, la codifica prestabilita risultava estremamente dettagliata riguardo le tipologie di correzione del dattiloscritto, il che non forniva, infine, dettagli interessanti e/o utili per l'analisi che ci si era preposte. D'altra parte, la codifica non prevedeva aspetti che sarebbero invece stati utili e pertinenti, come una annotazione che indicasse l'equivalenza tra le diverse forme del nome di uno stesso personaggio. Il caso di varianti grafiche è d'altronde estremamente rilevante nel corpus, vista la continua traslitterazione tra alfabeto arabo e latino, nonché la relativa instabilità della forma scritta della *darija* marocchina. Inoltre, a causa delle diverse forme dei testi, gli stessi elementi di codifica sono stati in alcuni casi usati in modi differenti (come le note, usate in un caso per le note d'autrice e nell'altro per quelle di editrice). Per quanto riguarda la seconda tipologia di edizione, anche qui il vocabolario delle annotazioni non è risultato completamente soddisfacente, e per i limiti di EVT e per i limiti della TEI. EVT essendo un software open source, lavorando sul codice probabilmente tali limiti si sarebbero potuti risolvere. Tuttavia, si tratta di un *workflow* pensato per essere portato avanti da persone senza una tale competenza informatica. Da una parte, si sono dovute fare delle scelte di tag obbligate per rispettare il vocabolario ottimizzato per EVT, mentre tag differenti sarebbero stati più adatti: ad esempio, si è dovuto marcare i termini in lingua straniera con il tag `<term>` seguito dall'attributo "ref" invece di `<foreign xml:lang>`, per poterne costruire una *list of entities*. Dall'altra parte, il vocabolario TEI in sé ha posto dei limiti non permettendo, ad esempio, di annidare i tag `<said who>` l'uno dentro l'altro, e quindi di marcare il meccanismo narrativo del racconto nel racconto, caratterizzato dalla presa di parola diretta di un personaggio all'interno del corpo del testo. L'espediente è nondimeno molto significativo nel corpus Chimenti, e si ispira non a caso al modello del racconto orientale come le *Mille e una notte*. In effetti, nei testi di Chimenti spesso troviamo raccolte che partono da una cornice narrativa per poi annidarvi all'interno ulteriori livelli di narrazione. Lo stesso racconto *Histoire du prince Moustapha et de la princesse Djamila*, oggetto della seconda edizione presentata, ne è un esempio. Il racconto è inserito ne *La veillée du harem*, raccolta che si apre con una cornice narrativa nella quale tutti i racconti che seguono sono iscritti. A sua volta, all'interno dei racconti sono presenti altri racconti narrati dai personaggi. Ancora, all'interno di questi racconti sono presenti dialoghi.

In maniera generale, è fondamentale decidere da subito un insieme di tag e un insieme che sia ragionevole e ben delimitato. Ciò ha due vantaggi diretti, e de un punto del lavoro umano e dell'agilità dei passaggi del workflow. Il primo vantaggio evidente è che un insieme di tag delimitato ed agile facilita l'apprendimento e la padronanza del sistema di codifica. Dall'altra, un sistema complesso di codifica si ripercuote sulla produzione dei file XML, rischiando di generare documenti estremamente complessi, difficili da leggere e manipolare. Nel primo workflow, il documento XML/TEI dell'edizione può essere manipolato solo tramite trasformazioni XSLT, il che fa perdere al file XML le sue caratteristiche user-friendly – già relative. Nel caso del secondo workflow, è stato scelto un insieme di tag limitati, il che ha permesso che il lavoro fosse portato avanti da una sola persona in modo sì più ridotto ma ordinato e producendo infine un documento XML/TEI ben leggibile e manipolabile.

In conclusione, è risultato evidente che la formalizzazione dei dati permette di far emergere alcuni aspetti oscurandone necessariamente altri, oltre che facendo perdere alcune dimensioni interpretative [20, 23]. È estremamente complesso (se non impossibile) rappresentare tutta la varietà del corpus con un unico sistema di codifica. Il prossimo obiettivo sarebbe dunque di mettere a punto un nuovo *workflow* che permetta di differenziare l'annotazione in base alle tipologie di testo, mantenendo le diverse rappresentazioni coerenti ed omogenee tra loro. L'annotazione strutturale dovrebbe essere separata dall'annotazione critica, a sua volta variabile in base alle necessità. Infine, un'ulteriore sfida è quella della interdisciplinarità. Trattandosi di un corpus ibrido, è importante che il lavoro presupponga un dialogo tra specialisti di ambiti diversi, anche geograficamente distanti. Un fitto scambio interdisciplinare è d'altronde pratica connaturata ai progetti di informatica umanistica, e in una misura inedita per il campo strettamente letterario [20, 21]. Tuttavia, nel pratico tale lavoro d'équipe non sempre è semplice da realizzare: i contesti di lavoro nei diversi lati del Mediterraneo non hanno accesso agli stessi strumenti, metodi ed infrastrutture, il che genera disparità a tutti i livelli del trattamento informatico dei dati.

5. CONCLUSIONI

Il plurilinguismo, l'interdisciplinarietà, la transnazionalità sono alcuni dei volti dello spazio mediterraneo, che se da una parte lo rendono un campo di studi fertile e proficuo, allo stesso tempo ne determinano i limiti e le difficoltà di azione. Fare rete in uno spazio simile vuol dire ancora oggi trovarsi davanti continue frontiere. Ad un livello metodologico, ancora le storie letterarie sono strettamente nazionali, così come rigidi sono i confini disciplinari, e i profili – di autori/autrici e di ricercatori/ricercatrici – che attraversano tali confini hanno ancora difficoltà ad essere integrati nel sistema. Ad un livello

scientifico, l'ibridismo e il *métissage* linguistici e culturali, se da un lato sono sempre più oggetto di studio nella ricerca, dall'altro pongono ancora diverse sfide, teoriche e pratiche, rispetto al modello di un monolinguismo occidentale. Ad un livello tecnico, è complesso rappresentare la pluralità mediante un sistema di rappresentazione formale. Se da una parte è evidente che per definire necessitiamo di categorie determinate e delimitate, dall'altra è fondamentale riflettere, oltre che sulle possibilità che tali categorizzazioni aprono, anche sui limiti che portano con sé.

BIBLIOGRAFIA

[1] Aghzal, M., M. A. E. Bouni, S. Driouech, e A. Mourhir. «Compact Transformer-based Language Models for the Moroccan Darija». In *7th IEEE Congress on Information Science and Technology (CiSt)*, 299–304. Agadir - Essaouira, Morocco, 2023. https://doi.org/10.1109/CiSt56084.2023.10409912.

[2] Bahri, Farid. *Tanger. Une histoire-monde du Maroc*. Paris: Les Éditions Bibliomonde, 2022.

[3] Bhabha, Homi K. *The location of culture*. London/New York: Routledge, 1994.

[4] Braudel, Fernand. *La Méditerranée et le monde méditerranéen à l'époque de Philippe II*. Paris: Armand Colin, 1949.

[5] Budor, Dominique, e Walter Geerts, (a cura di). *Le texte hybride*. Paris: Presses Sorbonne Nouvelle, 2004.

[6] Cederna, Camilla M. «Elisa Chimenti (Naples, 1883-Tanger, 1969): la traversée infinie de l'écriture entre les mondes, entre les langues». a cura di Elsa Chaarani, Laurence Denooz, e Sylvie Thiéblemont-Dollet, 529–44. Nancy: Presses universitaires de Lorraine, 2021.

[7] Cederna, Camilla M. «L'écriture d'exil d'Elisa Chimenti, une mosaïque des voix et de créations des femmes entre métissage et transgression». *Atlante. Revue d'Études romanes* 18 (2023a).

[8] Cederna, Camilla M. «Un projet international de recherche et d'édition pour la valorisation et la sauvegarde de l'œuvre d'Elisa Chimenti». *Atlante. Revue d'Études romanes* 18 (2023b).

[9] Chambers, Iain. *Mediterranean Crossings: The Politics of an Interrupted Modernity*. Durham, NC: Duke University Press, 2008.

[10] Chimenti, Elisa. *Anthologie*. Mohammedia/Casablanca: Éditions du Sirocco/Senso Unico Éditions, 2009.

[11] Deplano, Valeria. «Archivi d'Africa. Le carte dell'amministrazione coloniale in Italia e nei territori di nuova indipendenza». In *Archivi sul confine. Atti del convegno internazionale, Torino, Archivio di Stato, 6-7 dicembre 2017*, 41–52. Roma: Ministero per i beni e le attività culturali e per il turismo. Direzione generale archivi, 2019.

[12] Desideri, Ada, e Pauline Modolo. «Traitements informatiques des données d'un traitement de texte à une structuration XML TEI pour la création d'interfaces de consultation en HTML». *Atlante. Revue d'Études romanes* 18 (2023).

[13] Eichhorn, Kate. *The Archival Turn in Feminism: Outrage in Order*. Temple University Press, 2013.

[14] Fabre, Thierry, e Robert Ilbert. *Regards croisés sur la Méditerranée. «Les» représentations de la Méditerranée*. Paris: Maisonneuve & Larose, 2000.

[15] Foucault, Michel. *L'archéologie du savoir*. Paris: Gallimard, 1969.

[16] Hogan, Cristophe. «OCR for Minority Languages». In *1999 Symposium on Document Image Understanding Technology*. Annapolis: Maryland, 1999.

[17] Laplantine, François, e Alexis Nouss. *Le Métissage*. Paris: Flammarion, 1997.

[18] Matvejevic, Predrag. *Mediterraneo. Un nuovo breviario*. Milano: Garzanti, 1991.

[19] Miège, jean-Louis, Georges Bousquet, Jacques Denarnaud, e Florence Beaufre. *Tanger. Porte entre deux mondes*. Paris: ACR édition internationale, 1992.

[20] Moretti, Franco. *Falso movimento. La svolta quantitativa nello studio della letteratura*. Milano: Nottetempo, 2022.

[21] Moretti, Franco, *La letteratura vista da lontano*. Torino: Einaudi, 2005.

[22] Moussa, N.H., e A. Mourhir. «Named Entity Recognition in the Moroccan Dialect». In *7th IEEE Congress on Information Science and Technology (CiSt)*, 282–86. Agadir - Essaouira, Morocco, 2023. https://doi.org/10.1109/CiSt56084.2023.10409973.

[23] du Mouza, Cédric, Stéphane Lamassé, e Philippe Rygiel. «De l'histoire numérique à l'histoire données?» 42 (2023).

[24] Murrieta Flores, Patricia. «Language as cosmovisión/Cēmānāhuac. Some reflexions about computational research on the "Conquest of America" and the cruciality of decolonial praxis in the Digital Humanities». In *DH Benelux*. Bruxelles, Belgio, 2023.

[25] Nahli, O., N. Gugliotta, Giulia B., e B. Khlif. «Challenges and Progress in Constructing Arabic Dialect Corpora and Linguistic tools: A Focus on Moroccan and Tunisian Dialects». In *7th IEEE Congress on Information Science and Technology (CiSt)*, 293–98. Agadir - Essaouira, Morocco, 2023. https://doi.org/10.1109/CiSt56084.2023.10410009.

[26] Sapegno, Maria Serena. «Uno sguardo di genere su canone e tradizione». In *Dentro/ fuori, sopra/ sotto: critica femminista e canone letterario negli studi di italianistica*, a cura di Alessia Ronchetti e Maria Serena Sapegno. Ravenna: Longo Editore, 2007.

[27] Tamburini, Francesco. «Le istituzioni italiane di Tangeri (1926-1956): "quattro noci in una scatola", ovvero, mancati strumenti al servizio della diplomazia». *Africa: Rivista trimestrale di studi e documentazione dell'Istituto italiano per l'Africa e l'Oriente* 61, fasc. 3/4 (2006): 396–434.

[28] Vallarano Bianca. «“Oltre la lingua franca”. Il plurilinguismo mediterraneo di Elisa Chimenti». *Annali. Sezione romanza* 65, fasc. 1 (2023): 139–56.

[29] Windsor, Leah C. «Bias in Text Analysis for International Relations Research». *Global Studies Quarterly* 2, fasc. 3 (2022).

Mediterraneo verde fantastico: un repertorio digitale attraverso la collezione botanica di Ulisse Aldrovandi (1522-1605)

Sara Obbiso
Università di Bologna, Italia - sara.obbiso2@unibo.it

ABSTRACT

Il contributo si propone di presentare il progetto di un repertorio digitale di flora mediterranea di tipo magico, mostruoso e fantastico, tratto dalla collezione cinquecentesca del naturalista bolognese Ulisse Aldrovandi. Saranno raccolti alcuni esempi di piante ed erbe dotate di caratteristiche magiche, mostruose, mitologiche o alchemiche, prettamente incentrati sul territorio del Mediterraneo, origine di tutti gli studi classici sul mondo naturale, che serviranno a mostrare i legami tra le risorse, sia in termini di oggetti fisici provenienti dalla collezione botanica aldrovandiana, che delle diverse versioni digitalizzate degli stessi. Questi esempi costituiranno gli *items* e i dati di una collezione digitale costruita attraverso la piattaforma Omeka S, contenitore virtuale per la gestione, conservazione e valorizzazione del patrimonio culturale digitale, che avrà lo scopo di raccogliere e mettere in connessione non solo oggetti, luoghi, persone, conoscenze e fonti letterarie, che si rifanno ad un immaginario culturale comune, ma anche progetti, collezioni e risorse digitali già presenti online.

PAROLE CHIAVE

Natura; magia; cultura classica; collezione digitale; Ulisse Aldrovandi.

1. INTRODUZIONE

Il contributo proposto è il focus di un più ampio progetto di Dottorato in corso (2021-2024), dal titolo "Botaniche parallele", incentrato sull'analisi della botanica di tipo magico e fantastico a partire dalla collezione museale e libraria raccolta dal botanico e naturalista Ulisse Aldrovandi nel corso della sua vita (1522-1605) e ora conservata tra la Biblioteca Universitaria di Bologna e il Museo di Palazzo Poggi [6].

Egli, infatti, fu autore di trattazioni che oscillano tra lo scientifico e il sovrannaturale e in campo botanico si trovano spesso al confine con il mondo della criptobotanica o botanica fantastica, permeato di magia e simbolismo. Per questo motivo Ulisse Aldrovandi ha da sempre polarizzato molto l'interpretazione degli studiosi, che tendono da un lato ad esaltarlo come grande innovatore, dall'altro a demolirlo come latore di idee pseudoscientifiche. Sicuramente a lui è riservato un ruolo speciale nella storia della botanica, dal momento che allo stato attuale delle conoscenze, la sua raccolta di piante secche risulta la più ampia, la più ricca e la più diversificata fra quelle rinascimentali e il suo lascito costituisce un nucleo compatto e completo, che può fornire numerosi spunti anche per ricerche d'ambito storico, mitologico, floristico-ecologico e fitogeografico [8]. Infatti, l'enorme patrimonio di conoscenze ereditato da Aldrovandi, che ha esplorato ogni aspetto del mondo naturale con un metodo innovativo e moderno per l'epoca [17], nasconde numerosi esempi di quell'immaginario fantastico, ricco di aspetti magici, mostruosi, mitologici e simbolici, che affonda le radici nella classicità, permea la mentalità rinascimentale e si muove in maniera carsica nei secoli successivi fino alla contemporaneità, riaffiorando nei nomi popolari delle piante [5] e nelle tradizioni delle piccole comunità in territori di margine, in particolare nelle regioni meridionali dell'Europa e dell'Italia che si affacciano sul Mediterraneo, dove resistono con più forza credenze, superstizioni, rimedi terapeutici, pronostici e tecniche apotropaiche legate all'agricoltura che sono in continuità con la tradizione greco-romana e sono legate a concezioni come la simpatia e l'analogia [10]: ne sono esempio le erbe afrodisiache come l'abrotono, antiafrodisiache come l'agnocasto, apotropaiche come l'aglio, simpatetiche come il dittamo o legate a raccolte magiche e rituali come l'elleboro e la mandragola [9].

Proprio la botanica del Mediterraneo è, in effetti, il territorio su cui si fonda lo studio della botanica europea condotto da Ulisse Aldrovandi e dalla rete di protobotanici e naturalisti del XVI secolo con cui egli era in contatto, poiché terra in cui la civiltà classica si è sviluppata e dunque modello per eccellenza: testi come il *De materia medica* di Dioscoride, la *Naturalis Historia* di Plinio il Vecchio o l'erbario attribuito allo Pseudo-Apuleio, punti di riferimento per oltre quindici secoli, erano infatti incentrati principalmente sulla flora e la tradizione mediterranea.

La varietà tipologica è ampia, ma per nominare solo alcuni esempi che possiamo trovare tra le opere e i manoscritti del Fondo aldrovandiano occorre menzionare almeno: le erbe *lunarie* contenute negli erbari alchemici e legate a concezioni astrologiche [13, 14]; le piante mostruose, dipendenti da fenomeni come trasmutazioni, malformazioni, eventi prodigiosi, zoomorfia e antropomorfia, sincarpie e sinspermie, come i numerosi casi di agrumi deformi [2, 3: 663-715]; le piante

prodigiose come l'albero dell'isola El Hierro da cui piove copiosamente, che si trova presso la località chiamata volgarmente El Bajador, a Fuerteventura nelle Canarie [2: 35]; le piante mitiche, forse mai esistite realmente, come l'erba *moly* di cui parla Omero in *Odissea* X, 302–306, che in quest'opera veniva presentata come antidoto agli incantesimi di Circe e che va identificata probabilmente con la specie mediterranea dell'*Allium nigrum* L. [7]. L'esempio che probabilmente può racchiudere tutte le tipologie è, invece, quello della mandragola, erba mostruosa perché antropomorfa, alchemica poiché contenuta negli erbari alchemici dove è indicata per curare tutte le ferite e i problemi di infertilità, ma anche mitica e fantastica perché legata fin dall'antichità a molte leggende relative alla sua raccolta.

2. OBIETTIVI

Scopo del presente contributo è quello di partire dal focus specifico incentrato sulla botanica fantastica d'area mediterranea per presentare il cosiddetto *workflow di progetto* [16] del più ampio lavoro che lo contiene e che si estende anche ad altre aree geografiche in Europa e nel mondo da cui provenivano i campioni studiati da Aldrovandi: verrà, dunque, innanzitutto, tracciata l'analisi degli obiettivi, degli utenti e delle tipologie di contenuti trattati, mentre nei successivi paragrafi verrà fatto riferimento anche ai progetti simili che si intende prendere come riferimento, alla tipologia di dati, metadati e ontologie che si intende utilizzare nella descrizione delle risorse, alla piattaforma che farà da contenitore del progetto, alla sua usabilità e così via.

Scopo generale della ricerca, infatti, è quello di ricostruire le connessioni tra gli esempi di carattere botanico, i numerosi oggetti culturali museali e d'archivio dalla natura molteplice – opere a stampa, manoscritti, tavole acquerellate, xilografie – e i luoghi, le persone e le fonti letterarie, che attraverso gli esempi tratti della ricca collezione aldrovandiana permettano di esplorare il tema del rapporto tra l'uomo e il mondo naturale magico e fantastico. In questo modo si punta anche ad abbracciare un'ampia offerta di contenuti, che in un'ottica multidisciplinare possa attrarre un'utenza diversificata e trasversale.

Tra gli obiettivi principali, inoltre, vi è quello di stabilire quante più connessioni possibili con risorse utili già digitalizzate e presenti online, con il risultato sia di riutilizzare e valorizzare il materiale già esistente, ma anche di costruire un progetto paradigmatico per implementazioni future, che possa dunque non soltanto aiutare a ricostruire la storia di un contesto storico e culturale con le sue molteplici connessioni tra passato e futuro, ma anche offrire un modello scalabile per progetti simili calati in altre fasi temporali, altri contesti spaziali e culturali o a partire da collezioni museali e librarie dello stesso autore o di altri: tra gli sviluppi futuri del progetto potrà in questo modo essere valutata l'opportunità di replicare lo stesso lavoro in termini di spostamento del caso di studio dal regno vegetale agli altri regni del mondo naturale studiati da Aldrovandi, come il regno umano, zoologico, dei corpi celesti o dei minerali, anch'essi strettamente connessi a numerosi oggetti da collezione, tavole iconografiche e notizie afferenti al campo dalla magia, del fantastico e del mostruoso, come risulta evidente già da un veloce sguardo ai capitoli della *Monstrorum Historia* [3][1]. Da valutare, inoltre, sarà l'opportunità di integrare all'interno del progetto altri repertori di credenze e usanze in campo botanico già mappati, come quello ad opera di Emanuele Lelli, legato ai numerosi casi di continuità tra folklore antico e moderno [10].

3. STATO DELL'ARTE

Gli studi su Ulisse Aldrovandi hanno goduto nel corso dell'ultimo secolo di un'attenzione sempre maggiore, soprattutto a partire dalle celebrazioni per il terzo centenario della sua morte (1905-1907), di cui di recente è stata curata la digitalizzazione e la metadatazione di alcuni saggi e di estratti da pubblicazioni rare e di difficile accesso.

Nei primi anni 2000, in corrispondenza del quarto centenario dalla morte di Aldrovandi, si assiste a una nuova ondata di pubblicazioni e progetti, anche digitali, come quello coordinato dal Prof. Marco Beretta e fondato sulla creazione della piattaforma «Il Teatro della Natura di Ulisse Aldrovandi», che contiene un primo progetto solo in parte completato di digitalizzazione della documentazione aldrovandiana in possesso della Biblioteca Universitaria di Bologna: ne fanno parte il catalogo dei manoscritti, le immagini delle tavole acquerellate, parte dell'epistolario, il *Discorso naturale* e l'edizione digitale della biografia di Aldrovandi realizzata da Giovanni Fantuzzi[2]. Già in questa occasione erano stati sperimentati alcuni tentativi di catalogazione e archiviazione di testi, bibliografia e immagini, che necessitano però di essere riesaminati e aggiornati alla luce dei nuovi standard internazionali [11].

Nel 2004, per motivi sia di conservazione che di divulgazione, viene acquisito digitalmente l'intero corpus dell'erbario secco aldrovandiano, che consta di più di cinquemila campioni suddivisi in quindici volumi rilegati: il lavoro, corredato

[1] Cfr. l'indice dell'opera digitalizzata accessibile al seguente link: https://historica.unibo.it/handle/20.500.14008/78076.
[2] Il progetto è raggiungibile al link http://aldrovandi.dfc.unibo.it/, da cui si accede anche al link dell'archivio online http://moro.imss.fi.it/aldrovandi/.

dall'apparato critico a cura di Adriano Soldano, è stato pubblicato all'interno del portale del Sistema Museale di Ateneo[3].
Nel triennio 2005-2007 prese, inoltre, avvio il progetto di censimento, digitalizzazione in alta risoluzione e catalogazione delle circa quattromila matrici xilografiche di Ulisse Aldrovandi, inizialmente fruibili ad uso soltanto interno, ma attualmente disponibili online in parte sul sito del "Catalogo regionale del Patrimonio culturale dell'Emilia-Romagna"[4], dove si trovano 1820 incisioni, mentre dal 2020 sul sito del Catalogo online del Sistema Museale d'Ateneo, al momento nel numero di circa duecento risultati corredati da una ricca scheda e dall'immagine dell'oggetto accompagnata dal suo IIIF Manifest[5].
Questi primi progetti sono stati poi seguiti da numerosi altri, nati grazie alle attività che l'Università di Bologna ha promosso per le celebrazioni del quinto centenario della nascita di Aldrovandi e in particolare in occasione della mostra temporanea «L'Altro Rinascimento», che ha consentito di lavorare al recupero e allo studio di materiale inedito, nonché alla digitalizzazione delle opere e dei manoscritti di Aldrovandi, materiale che ora è messo a disposizione online attraverso la piattaforma AMSHistorica[6]. Su questi e altri output realizzati e presentati in occasione della mostra si sta ora lavorando con il proposito di una digitalizzazione basata sui Linked Open Data, che vada nella direzione dei principi FAIR dell'Open Science: lo scopo è quello di realizzare il cosiddetto *digital twin* della mostra, una replica digitale che possa consentire di organizzare e rendere accessibile sul web il percorso espositivo attraverso l'esplorazione virtuale di tutti gli *items* multimediali prodotti, nonché di analizzare e visualizzare dati e metadati che confluiranno all'interno di un'apposita *digital library online*. È questo l'obiettivo già parzialmente realizzato di uno dei nove sotto-progetti tematici del Progetto CHANGES[7], finanziato dal PNRR e dedicato proprio all'utilizzo di tecnologie virtuali per la valorizzazione del patrimonio culturale di musei e collezioni d'arte [4]. Infine, è attualmente in corso il progetto di Edizione Nazionale delle Opere di Ulisse Aldrovandi, basata su una piattaforma digitale integrata: anche questo progetto prevede, infatti, di includere non solo le opere ma anche gli oggetti della collezione aldrovandiana, come l'erbario secco, le xilografie e il catalogo degli oggetti museali di Palazzo Poggi e di altri musei italiani cui si fa esplicito riferimento nelle opere [11].
Infine, bisogna tenere in considerazione i numerosi collegamenti con erbari online che contengono casi simili e altri oggetti relativi alla cultura botanica fantastica, preservati all'interno di importanti archivi e collezioni digitali: un esempio è sicuramente l'archivio digitale PHAIDRA dell'Università di Padova [15], che ha già messo a disposizione una collezione di erbari illustrati, immagini botaniche e tavole risalenti ai secoli dal XV al XX. All'interno di questa collezione si trova un esemplare di erbario manoscritto e illustrato attribuito allo Pseudo-Apuleio e risalente all'ultimo quarto del secolo XV, che tra le piante officinali riporta una versione fantastica ma piuttosto realistica di una mandragola (c. 148*v*), che sarà interessante confrontare con le mandragole aldrovandiane. Altri esempi utili e interessanti sono quelli relativi ad alcuni erbari della tradizione alchemica già in parte o del tutto metadatati e digitalizzati da parte di diverse istituzioni all'interno delle loro biblioteche digitali: si tratta, ad esempio, del caso della *Bodleian Library* di Oxford con il ms. Canon. Misc. 408, della *Wellcome Historical Medical Library* di Londra con i mss. 261, 334 e 337 o della *BnF* con i mss. lat 17848, lat. 17844 ed hébr. 1199. Altri utili modelli di confronto e di analisi saranno, infine, i lavori di digitalizzazione dei campioni di erbari essiccati, come quello dedicato all'*Herbarium Horti Botanici Pisani*[8].
Con tutti i progetti finora elencati e con altri simili ci si vorrà, dunque, porre proficuamente in dialogo, con lo scopo di riunire e valorizzare le risorse già esistenti e offrire nuovi percorsi di lettura, sia all'interno dello specifico patrimonio aldrovandiano che riguardo il più ampio rapporto che intercorre tra l'essere umano e il mondo naturale in termini sociali e culturali.

4. METODOLOGIE

Gli esempi di botanica fantastica tratti dalla collezione di Ulisse Aldrovandi saranno raccolti nella forma di un repertorio digitale costruito attraverso la piattaforma Omeka S[9], software che funziona da contenitore virtuale per la conservazione, la descrizione e la valorizzazione di piccole collezioni digitali ed è stato creato appositamente per le istituzioni interessate a collegare il proprio patrimonio culturale digitale con altre risorse già presenti online [12].

[3] Il progetto è consultabile al seguente link: http://botanica.sma.unibo.it/.
[4] Le schede sono consultabili al seguente link: https://bbcc.regione.emilia-romagna.it/pater/search.do?value%28NCTA%29=BO057&option%28NCTA%29=strict&type=mi&fakesearch=Incisioni+collegate
[5] Il catalogo del Sistema Museale di Ateneo è pubblicato al seguente indirizzo: https://catalogo.sma.unibo.it/it/29/ricerca/iccd/?search=museo+di+palazzo+poggi&paginate_pageNum=1&facet%5B0%5D=OGTD_ss%3A%22matrice+xilografica%22&facet%5B1%5D=SGTI_ss%3A%22Botanica%22.
[6] https://historica.unibo.it/cris/fonds/fonds02019
[7] La sigla CHANGES sta per "Cultural Heritage Active Innovation For Next-Gen Sustainable Society".
[8] https://erbario.unipi.it/it
[9] L'istanza di Omeka S è installata sul server del Laboratorio FrameLAB, che fornisce il supporto tecnico al progetto e sarà resa pubblica alla conclusione del percorso di Dottorato, contemporaneamente alla pubblicazione della tesi.

All'interno di ogni istanza di Omeka gli oggetti, detti *items*, possono essere accompagnati dai loro media, visualizzati attraverso appositi strumenti di visualizzazione come *Mirador*, possono essere raggruppati in collezioni dette *items sets* in base alle loro caratteristiche e possono essere descritti grazie ad un insieme di metadati definiti a priori all'interno di un specifico modello di risorsa, cioè un insieme di proprietà predefinite rispondenti ad un preciso linguaggio standardizzato chiamato ontologia.
In questo caso verranno stabilite relazioni tra risorse di tipo testuale e immagini, che riguardano principalmente documenti d'archivio, opere a stampa, descrizioni catalografiche e fotografie di oggetti eterogenei. Le entità individuate come fondamentali per la modellazione delle relazioni tra questo tipo di risorse sono: la *pianta*, la *fonte* aldrovandiana da cui essa viene estrapolata a partire dallo spoglio delle opere e dei manoscritti, la *notizia* che caratterizza quella pianta in senso fantastico, la *citazione* cui Aldrovandi fa riferimento per accreditare la notizia e che solitamente si trova a margine del testo sia manoscritto che a stampa, infine l'*autore* di questa citazione. Altre entità potranno essere introdotte e collegate in futuro per arricchire il progetto e portarlo ad un maggior livello di dettaglio e complessità, come ad esempio un'istanza di tipo testuale, che descriva la ricetta o il consiglio medico in cui si trova citata quella determinata erba o pianta all'interno di altre opere aldrovandiane più specifiche come l'*Antidotarii Bononiensis* [1].
Prendendo come esempio il modello di risorsa dell'entità principale, la *pianta*, esso conterrà sicuramente alcuni metadati come: un titolo, riferito al nome con cui la pianta si trova più frequentemente all'interno delle opere o dei manoscritti di Aldrovandi e che costituisce anche il nome dell'*item*; un identificativo riferito alla tassonomia scientifica, identificato attraverso la letteratura analizzata e definito come risorsa digitale grazie al collegamento con il progetto *Plants of the World Online*, che di ogni pianta ha fornito non solo il nome e la descrizione, ma anche la distribuzione geografica[10]; un titolo alternativo, che raccolga gli altri nomi con cui ci si riferisce nelle fonti alla stessa pianta e tenga quindi conto della proliferazione dei nomi in campo botanico ancora tipica dell'epoca premoderna; una descrizione, che raccolga in forma testuale le informazioni storiche e culturali più importanti; un indice dei contenuti, che colleghi la pianta alle altre risorse digitalizzate della collezione aldrovandiana, come le tavole acquerellate, le matrici xilografiche e l'esemplare agglutinato della stessa conservato nell'erbario secco; una citazione bibliografica, il cui valore sarà un *item* di Omeka creato attraverso il modello di risorsa relativo alle citazioni; la copertura territoriale, che servirà a localizzare geograficamente la pianta sulla base di quanto indicato da progetti autorevoli come *Plants of the World Online*, in modo tale che attraverso questo metadato sia possibile isolare gli oggetti provenienti specificamente da alcune aree come quella mediterranea e partendo da questa selezione condurre un'analisi separata.
Per quanto riguarda la scelta delle ontologie più adatte per i metadati che comporranno ciascun modello di risorsa, verranno innanzitutto utilizzate quelle offerte di *default* con l'installazione dell'istanza di Omeka: quindi, ad esempio, DublinCore per gli oggetti eterogenei come le piante, FOAF per gli autori e BIBO per le risorse bibliografiche.
Ciascun *item*, se ritenuto utile, potrà far parte di una collezione, ma sicuramente le piante saranno divise e raggruppate in quattro collezioni in base alla loro natura di piante mitiche e fantastiche, mostruose, alchemiche o esotiche, oppure in più di una tra queste se si tratterà di una pianta dalla natura ibrida, come nel caso della mandragola.
Si prevede, infine, di offrire all'utente finale una navigazione delle risorse attraverso un sito web costruito grazie alla funzionalità di Omeka di agire anche come un CMS e lavorare su dei *template* preimpostati o su quelli personalizzati e condivisi dalla *community*. Una ricerca degli *items* tramite la tipologia a faccette sarà implementabile grazie ad un modulo specifico di Omeka S, sulla base di modelli già esistenti e consolidati, come l'esperienza che della piattaforma si è fatta quando è stata implementata la *Digital Library* del Dipartimento FICLIT dell'Università di Bologna, come strumento di ricerca, didattica e terza missione [7], o in progetti simili, come nel caso della *Digital Library* della Biblioteca Classense di Ravenna[11].

5. RISULTATI OTTENUTI O ATTESI

Sono state diverse le occasioni accademiche in cui è stato possibile presentare alla comunità di ricercatori i primi risultati dell'indagine su segmenti specifici della botanica fantastica aldrovandiana, come il convegno *Herbaria* promosso dall'Università di Padova o il Convegno SISS di Napoli del 2023, dove nello specifico sono state analizzate le piante appartenenti alla tradizione alchemica. Per quanto riguarda, invece, i risultati digitali, essi sono stati sperimentati attraverso progetti strutturati in maniera molto simile, come quello a cui si è lavorato nel corso del 2023 riguardante la realizzazione di una collezione digitale costruita con Omeka S, che contenesse le vignette metadatate della Quarantana, cioè l'edizione

[10] https://powo.science.kew.org/
[11] Alla definizione del progetto per la definizione di una Digital Library per la Biblioteca Classense si è collaborato insieme al suo staff nel corso dell'a.a. 2021-2022.

illustrata dei *Promessi Sposi* del 1840[12]. Questo strumento, proposto sia per la ricerca che per la didattica scolastica e universitaria, è stato presentato in occasione di un'attività di Terza Missione rivolta alla formazione di alcuni insegnanti delle Scuole Superiori. Entro il mese di maggio 2024 si prevede di poter realizzare un prototipo simile incentrato questa volta sulla collezione botanica aldrovandiana, che contenga prima di tutto le immagini, le risorse e i dati di erbe e piante oggetto della già menzionata focalizzazione in area mediterranea, a partire dai casi più emblematici già in parte citati.
Il risultato finale atteso è quello di creare una solida base relazionale costituita dalla rete di risorse finora descritta, a partire dalla quale si possa in seguito lavorare a sviluppi ulteriori del progetto.

6. RINGRAZIAMENTI

Si ringrazia lo staff dei laboratori FrameLAB del Dipartimento di Beni Culturali di Ravenna e ADLAB del Dipartimento di Filologia Classica e Italianistica di Bologna per il supporto tecnico nell'implementazione dei progetti relativi alle collezioni digitali con Omeka S; si ringraziano la ricercatrice Noemi Di Tommaso e il team di ricercatori del Dipartimento di Storia della Scienza per la collaborazione alla realizzazione degli output digitali della mostra «L'Altro Rinascimento» e per le consulenze specifiche sui contenuti della stessa mostra; si ringrazia, infine, la Biblioteca Universitaria per il materiale inedito messo a disposizione.

BIBLIOGRAFIA

[1] Aldrovandi, Ulisse. *Antidotarii Bononiensis, siue de vsitata ratione componendorum, miscendorumque medicamentorum, epitome*. Bononiae: Giovanni Rossi, 1574.

[2] Aldrovandi, Ulisse. *Dendrologiae naturalis scilicet arborum historiae libri duo sylua glandaria, acinosumq. pomarium vbi eruditiones omnium generum vna cum botanicis doctrinis ingenia quaecunque non parum iuuant, et oblectant. Ouidius Montalbanus*. Bononiae: ex typographia Ferroniana, 1667.

[3] Aldrovandi, Ulisse. *Monstrorum historia cum Paralipomenis historiae omnium animalium. Bartholomaeus Ambrosinus ... labore, et studio volumen composuit. Marcus Antonius Bernia in lucem edidit. Proprijs sumptibus ... cum indice copiosissimo*. Bononiae: Marco Antonio Bernia, 1642.

[4] Balzani, Roberto, Sebastian Barzaghi, Gabriele Bitelli, e Federica Bonifazi. «"Saving Temporary Exhibitions in Virtual Environments: The Digital Renaissance of Ulisse Aldrovandi – Acquisition and Digitisation of Cultural Heritage Objects"». *Digital Applications in Archaeology and Cultural Heritage* 32 (marzo 2024). https://doi.org/10.1016/j.daach.2023.e00309.

[5] Beccaria, Gian Luigi. *I nomi del mondo: santi, demoni, folletti e le parole perdute*. Torino: Einaudi, 1995.

[6] Biancastella, Antonino. *L'Erbario di Ulisse Aldrovandi: natura, arte e scienza in un tesoro del Rinascimento*. A cura di Biancastella Antonino, Andrea Ubrizsy Savoia, e Alessandro Tosi. Milano: Federico Motta, 2003.

[7] Bonora, Paolo, Lucia Giagnolini, Alessandra Di Tella, e Francesca Tomasi. «Digital Library di dipartimento: da collettore di risorse digitali a strumento per la ricerca, la didattica e la terza missione». *Bibliothecae.it* 12, fasc. 2 (2023): 444–63.

[8] Buldrini, Fabrizio, Alessandro Alessandrini, Umberto Mossetti, Giovanna Pezzi, e Juri Nascimbene. «L'erbario di Ulisse Aldrovandi: attualità di una collezione rinascimentale di piante secche». *Aldrovandiana Historical Studies in Natural History* 2, fasc. 1 (2023): 7–34. https://doi.org/10.30682/aldro2301a.

[9] Lelli, Emanuele. *Folklore antico e moderno: una proposta di ricerca sulla cultura popolare greca e romana*. Pisa-Roma: Serra, 2014.

[10] Lelli, Emanuele, e Luca Aprile. «Botanica antica e moderna tra folklore e scienza». In *Appunti romani di filologia: studi e comunicazioni di filologia, linguistica e letteratura greca e latina*, 67–96. Pisa: Fabrizio Serra Editore, 2019.

[11] Redazione. «Il progetto di Edizione Nazionale». *Aldrovandiana. Historical Studies in Natural History* 1, fasc. 1 (2022): 95–116.

[12] Salarelli, Alberto. «Gestire piccole collezioni digitali con Omeka». *Bibliothecae.it* 5, fasc. 2 (2016): 177–200. https://doi.org/10.6092/ISSN.2283-9364/6393.

[13] Segre Rutz, Vera. «Capitulum de arbore borissa – Le piante della Luna». In *Florilegium: scritti di storia dell'arte in onore di Carlo Bertelli*, 124–29. Milano: Electa, 1995.

[14] Segre Rutz, Vera. *Il giardino magico degli alchimisti: un erbario illustrato trecentesco della Biblioteca universitaria di Pavia e la sua tradizione.* Milano: Il polifilo, 2000.

[15] Tallandini, Laura, Lorisa Andreoli, Elena Bianchi, Linda Cappellato, Yuri Carrer, Gianluca Drago, Giulio Turetta, e Antonietta Zane. «Phaidra, un archivio digitale FAIR per la disseminazione e l'accesso integrato a testi, testimonianze, immagini e storie del patrimonio culturale». *DigItalia* 14, fasc. 1 (2019): 147–57.

[16] Tomasi, Francesca. *Organizzare la conoscenza: digital humanities e web semantico*. Milano: Editrice bibliografica, 2022.

[17] Tugnoli Pattaro, Sandra. *Metodo e sistema delle scienze nel pensiero di Ulisse Aldrovandi*. Bologna: CLUEB, 1981.

[12] Gli articoli derivati da questi lavori e interventi sono in corso di pubblicazione, prevista nel corso del 2024.

Per la digitalizzazione del patrimonio linguistico e culturale italiano in Egitto: I periodici italiani (1892- 1940)

Wafaa El Beih
Università di Helwan, Cairo, Egitto - Wafaa_elbeih@arts.helwan.edu.eg

ABSTRACT
L'intervento attuale vuole presentare gli sforzi dedicati alla conservazione e alla digitalizzazione del patrimonio culturale e linguistico italiano in Egitto, in particolare la prima fase del progetto della digitalizzazione dei periodici pubblicati in Egitto tra il 1892 e il 1940, compiuta entro il 2022. L'intervento, inquadrando l'importanza della stampa periodica italiana, come fonte indispensabile per ricostruire le vicende e le attività della comunità italiana d'Egitto, mette in luce le diverse fasi del progetto, gli esiti della prima fase e le iniziative e le scoperte nate a seguito di questi.

PAROLE CHIAVE
Digitalizzazione; periodici; stampa; patrimonio culturale; giornalismo.

1. INTRODUZIONE
Come ha giustamente sottolineato Umberto Rizzitano, la stampa periodica italiana costituisce una fonte indispensabile per ricostruire le vicende legate alla storia della colonia italiana d'Egitto, e tutta la storia del Paese in uno dei momenti più intensi della sua rinascita politica, economica e culturale [2]. Il giornalismo italiano iniziò in Egitto nel 1845 con «Lo Spettatore Egiziano» di Alessandria, un bisettimanale che fu pubblicato per oltre un quindicennio e si rivolgeva ad una vasta schiera di lettori italiani ed italofoni assolvendo le funzioni di Gazzetta Ufficiale del governo. Al giornale, fondato dall'avvocato Guido Leoncavallo, seguirono tanti altri come «Il Manifesto Giornaliero», «Il Progresso d'Egitto», «La Trombetta», «Il Giornale Marittimo», «L'Avvenire d'Egitto», «L'Eco d'Egitto», «Il Nilo», «Il Giornale d'Egitto» [2: 129-154].
Il quotidiano italiano che ebbe una vita più lunga è «L'imparziale» fondato al Cairo nel 1892 da Emilio Arus e fuso poi nel 1930 con il «Messaggero Egiziano» di Alessandria prendendo il nome di «Il Giornale d'Oriente» che rimase in vita fino al 1940. Il giornale, acquistato dal Fascio locale, e diretto da Giuseppe Galassi, vantava, durante i suoi primi mesi, una tiratura di 7.000 copie, ridotte poi a 4.000. Aveva sei pagine (arrivavano in qualche caso a dodici) ed era considerato uno dei migliori giornali dell'Egitto e del Medio Oriente. Dedicava attenzione alla politica internazionale, alla cronaca egiziana, locale ed italiana, quanto allo sport, alla moda, agli spettacoli. Durante il ventennio fascista esso sostenne attivamente le politiche fasciste[1].
Ma già all'epoca del suo articolo (1956), Rizzitano notò che c'erano pochissime informazioni sulla stampa, e sulla produzione editoriale italiana in Egitto, perché il materiale era stato in gran parte perso, ed era quindi difficile trovare i primi giornali nelle biblioteche pubbliche. A partire da questa preoccupazione, si è resa necessaria un'analisi approfondita della situazione attuale delle collezioni sopravvissute che costituiscono "una testimonianza unica o eccezionale di una tradizione culturale, secondo il quarto criterio delle linee guida della Convenzione per la Conservazione del Patrimonio Mondiale Culturale e Naturale, emanata dall'UNESCO nel 1972. Con la considerazione che questi giornali rappresentano effettivamente una parte del patrimonio materiale e immateriale dell'Italia e dell'Egitto, sia linguistico che culturale, è stato elaborato il progetto per la digitalizzazione di tali periodici italo-egiziani. Già nel 2019, sono nate le discussioni sul progetto

[1] Da sottolineare che la propaganda fascista si innestò, in Egitto, in un mondo coloniale già particolarmente avvezzo alla presenza di fogli italiani che si rivolgevano – con alterni successi – sia alla comunità di media e piccola borghesia che ruotava intorno ai consolati, sia alle masse lavoratrici. Nel corso degli anni Venti, i fascisti riuscirono a controllare e a far divenire strumento di propaganda *L'Imparziale*, il periodico che da trent'anni si era attestato come punto di riferimento per i membri della comunità italiana. I risultati ottenuti in questo campo, anche grazie al sostegno economico fornito dal governo a molte di queste testate, furono assolutamente ragguardevoli, poiché furono numerosissimi i giornali italiani in tutto il mondo che, con diverse modalità, si avvicinarono, come accadde con L'Imparziale, alle posizioni del regime fascista. Negli anni Trenta i fascisti tentarono progressivamente di modernizzare la propaganda nei confronti dei connazionali all'estero, distinguendosi nettamente dall'esperienza precedente. Allo spirito di controllo e guida delle comunità all'estero si sostituì una vera e propria volontà di conquista delle masse italiane in terra straniera, accompagnata dal tentativo di consolidare l'ideologia fascista al di fuori dei confini nazionali. *Il Giornale d'Oriente* dedicava una sezione per le notizie del Fascio, del Gruppo Sportivo Littorio e delle attività culturali del Circolo Italia-Dopolavoro. Insieme a quella sezione, c'erano di solito la cronaca sulle attività delle organizzazioni giovanili, in particolare le gite compiute dai giovani e dalle giovani fascisti al Cairo e ad Alessandria.

sia nella mia sede, l'Università di Helwan, istituendo un Centro di Umanistica Digitale per la Conservazione del Patrimonio Linguistico e Culturale in Egitto, sia nell'Istituto Italiano di Cultura, e alla fine nell'Ambasciata d'Italia.
Nel periodo tra il 2019 e il 2021, sono state avviate le ricerche sulle varie collezioni dei periodici italiani in Egitto, esaminando il loro stato di conservazione, e la miglior soluzione che possa garantire un accesso globale ad esse, abbattendo le tradizionali barriere dovute a distanze geografiche, a condizioni economiche o politiche. Un'indagine preliminare ha evidenziato tre collezioni di periodici italiani: la prima di proprietà del Centro Archeologico Italiano, affiliato all'Istituto Italiano di Cultura, la seconda si trova nella Biblioteca Nazionale Egiziana (Dār Al-Kutub), mentre la terza è nella Biblioteca Comunale di Alessandria. Le prime due collezioni abbracciano le testate: «L'Imparziale», «Il Giornale d'Oriente» e «Messaggero Egiziano», con annate diverse, mentre la terza, oltre a queste già citate, ne vanta varie altre, edite fino agli anni Cinquanta del secolo scorso.

2. PROCESSO DI ESECUZIONE

Nel periodo da gennaio del 2021 fino a settembre dello stesso anno, è stata individuata la prima collezione da digitalizzare, ossia quella dell'emeroteca del Centro Archeologico Italiano, sulla base delle agevolazioni concesse dall'Ambasciata d'Italia, in base alla notevole rilevanza del progetto. A questa fase è seguita quella relativa alla costituzione del gruppo di lavoro, alla distribuzione dei compiti e delle responsabilità in base ai requisiti e alle fasi del processo di digitalizzazione e agli output di ciascuna attività[2].
In seguito, sono stati discussi gli strumenti di acquisizione digitale adatti agli esemplari da digitalizzare e appropriati agli obiettivi del progetto, tenendo in considerazione che alcuni degli strumenti automatici che sono stati progettati per la digitalizzazione di massa potrebbero non essere appropriati per digitalizzare materiali rari e fragili che corrono il rischio di subire danni durante la procedura. Non solo, ma la qualità dell'immagine, la risoluzione, la profondità di colore e l'illuminazione sono parametri che dovrebbero essere decisi tenendo conto degli standard specifici largamente accettati per il tipo di materiale in questione, sulla base dei requisiti della presentazione e dell'uso previsti per gli oggetti digitali.
Considerando lo stato della collezione, e il bilancio disponibile, è stato deciso di affidare l'incarico dell'elaborazione della fotografia digitale a una società esterna, una tra le aziende che lavorano da anni nel campo della digitalizzazione delle risorse del patrimonio culturale, ed è insieme l'agente di una delle più grandi aziende internazionali produttrici di *overhead scanner* (vd. Fig. 1). In questa fase, sono state identificate le specifiche tecniche per i formati dei file digitali frutto dal processo di trasformazione digitale, sulla base di standard internazionali e sulle pratiche adottate dai progetti simili; oltre a ciò, è stato stabilito il sistema di denominazione che verrà utilizzato per nominare i file e le cartelle in cui verranno archiviati le pagine scannerizzate.
Il progetto è stato lanciato ufficialmente nel settembre del 2021, mentre il lavoro è cominciato il 16 gennaio 2022 e concluso il 6 agosto dello stesso anno. Entro quel periodo, il gruppo di lavoro si è impegnato a: Individuare e recuperare negli scaffali i volumi, esaminarli e descriverne le condizioni fisiche; Eseguire semplici lavori di restauro sulle pagine strappate, o piegate, a causa di una impropria conservazione; Contare accuratamente il numero di pagine di ogni numero, all'interno di ogni volume; elaborare dei record che registrano in dettaglio i calcoli; inserire i dati in un apposito file Excel per facilitare il processo di organizzazione automatica dei numeri e dei doppioni; Identificare i problemi di indicizzazione derivanti dalla presenza di numeri duplicati o errori di enumerazione; Scannerizzare le varie testate; Rivedere e regolare la qualità delle immagini digitali e memorizzarle su un apposito server; Riscannerizzare i file non conformi alle specifiche approvate dal progetto (vd. Fig. 2).

3. RISULTATI

Alla fine di questa fase del progetto, è stata elaborata una copia digitale dell'emeroteca del CAI che comprende 130 volumi di varie dimensioni, inclusi 20 doppioni. È stata creata una copia digitale di 110 volumi pubblicati tra gli anni 1892 e 1940, per un totale di 56 anni di giornalismo italiano in Egitto, 71.614 pagine, e uno spazio di archiviazione che supera 17 terabyte. (vd. Tab. 1)
Il materiale digitalizzato, insieme ad un database manuale (i calcoli dei numeri delle varie testate e delle cui pagine) e digitale (un file Excel per ogni testata e uno che raccoglie l'insieme), è conservato su un server di proprietà dell'Istituto Italiano di Cultura (vd. Figg. 3, 4, 5, 6). Questi esiti sono stati già annunciati in un Convegno internazionale sulla

[2] I membri del gruppo di lavoro variano tra specialisti in lingua e letteratura italiana e in tecniche di digitalizzazione e di trasformazione digitale: la sottoscritta, Wafaa Abdel Raouf El Beih, Direttore di ricerca e Ordinario di Letteratura Italiana Moderna e Contemporanea presso l'Università di Helwan, il Prof. Emad Issa Saleh (esperto di biblioteche digitali), e capo del Dipartimento di Scienza dell'Informazione presso la medesima sede, la Dott.ssa Marwa Salem (Manuscritti e restauro), il Dott. Amr Yahya (Archiviazione e conservazione del patrimonio culturale), il Dott. Tarek Hussein (Italianistica).

rivoluzione digitale tenutosi a Hurgada alla fine dell'anno scorso [*Lo sviluppo delle scienze umanistiche nell'era della digitalizzazione e della nuova Repubblica*, I Convegno internazionale della Facoltà di Lettere, Università di Kafr El Sheikh, Hurgada, 26- 29 ottobre 2022], e sono usciti negli atti dello stesso Convegno [1: 84-103]. La seconda fase per la medesima collezione coinvolge la British Library per una rielaborazione del materiale digitalizzato in modo che sia accessibile online ai ricercatori[3].

ANNO	PERIODICO	N. DI PAGINE	NOTE
1892	L'IMPARZIALE	1040	Un anno intero
1895	L'IMPARZIALE	1210	//
1897	L'IMPARZIALE	1216	//
1898	L'IMPARZIALE	1211	//
1900	L'IMPARZIALE	1212	//
1902	L'IMPARZIALE	1224	//
1904	L'IMPARZIALE	1212	//
1905	L'IMPARZIALE	1226	//
1906	L'IMPARZIALE	1208	//
1908	L'IMPARZIALE	1200	//
1909	L'IMPARZIALE	1218	//
1911	L'IMPARZIALE	1168	//
1912	L'IMPARZIALE	1222	//
1913	L'IMPARZIALE	1236	//
1914	L'IMPARZIALE	1282	//
1915	L'IMPARZIALE	1433	//
1917	L'IMPARZIALE	1322	//
1918	L'IMPARZIALE	1238	//
1920	L'IMPARZIALE	810	Da gennaio a settembre
1921	L'IMPARZIALE	1230	Un anno intero
1922	L'IMPARZIALE	600	Da gennaio a giugno
1923	L'IMPARZIALE	1364	Un anno intero
1924	L'IMPARZIALE	1258	//
1925	L'IMPARZIALE	1255	//
1926	L'IMPARZIALE	1207	//
1927	L'IMPARZIALE	1262	//
1928	L'IMPARZIALE	1446	//
1929	L'IMPARZIALE	1336	Da aprile a dicembre
1930	L'IMPARZIALE	448	Da gennaio a marzo
1926	MESSAGGERO EGIZIANO	614	Da luglio a dicembre
1927	MESSAGGERO EGIZIANO	1250	Un anno intero
1928	MESSAGGERO EGIZIANO	1454	//
1929	MESSAGGERO EGIZIANO	788	Da luglio a dicembre
1930	MESSAGGERO EGIZIANO	450	Da gennaio a marzo
1930	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	1389	Da aprile a dicembre
1931	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	1388	Da aprile a dicembre
1932	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	1930	Un anno intero
1933	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	1964	//
1934	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	1916	//
1935	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	1935	//
1936	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	1618	Da gennaio a settembre
1937	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	2144	Un anno intero
1938	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	2062	//
1939	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	1774	//
1940	IL GIORNALE D'ORIENTE (L'IMPARZIALE)	654	Da gennaio ai primi numeri di giugno
1930	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	924	Da luglio a dicembre
1931	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	1304	Da gennaio a giugno- da ottobre a dicembre
1932	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	1450	Da gennaio a giugno- Due numeri di luglio- Un solo numero di agosto- Da ottobre a dicembre
1933	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	1432	Da gennaio a giugno- da ottobre a dicembre
1934	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	968	Da gennaio a giugno
1935	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	1474	Da gennaio a settembre
1936	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	2174	Un anno intero
1937	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	1592	Da aprile a dicembre
1938	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	1044	Da luglio a dicembre
1939	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	1338	Da gennaio a giugno- da ottobre a dicembre
1940	IL GIORNALE D'ORIENTE (MESSAGGERO EGIZIANO)	290	Da aprile a giugno
	TOTALE	**71614**	

Tabella 1. Versione digitale della Collezione del CAI

[3] Riguardo ciò mi riferisco alla notizia pubblicata qui https://blogs.bl.uk/endangeredarchives/2023/08/stampa-migrante-a-window-on-multi-lingual-egypt.html.

Figura 1. L'overhead scanner dentro il Centro Archeologico Italiano

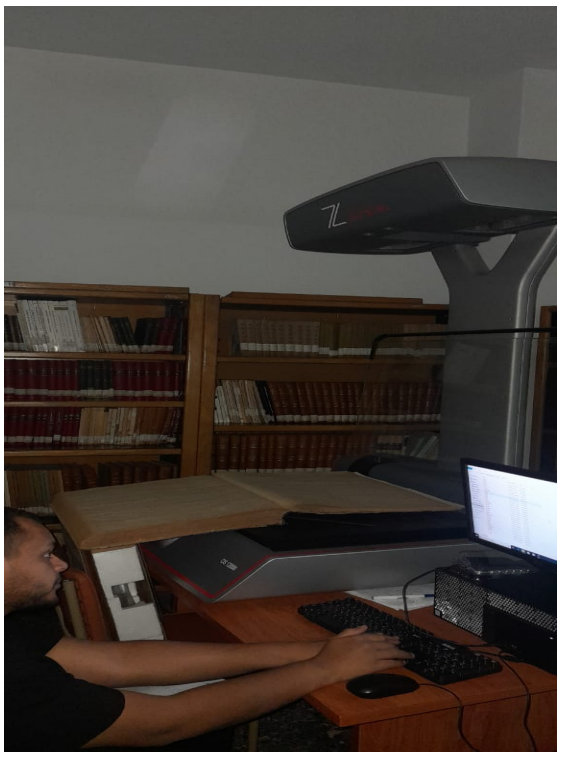


Figura 2. Processo di scannerizzazione, da 16 gennaio a 6 agosto 2022.

L'IMPARZIALE

DA PELLOUX A MUSSOLINI

Figura 3. Copia scannerizzata formato jpg dell'«Imparziale» del 9 marzo 1924.

IL GIORNALE D'ORIENTE

ALL'ACCADEMIA D'ITALIA

Il Giappone ha scatenato l'offensiva in Manciuria

contro il comandante cinese ribelle

I debiti di guerra

Quali sorprese riserba quest'anno il GIORNALE D'ORIENTE ai suoi abbonati 1933?

Figura 4. Copia scannerizzata bianco e nero formato jpg di «Il Giornale d'Oriente/Imparziale» del 2 dicembre del 1932.

Figura 5. Copia scannerizzata formato jpg di «Il Giornale d'Oriente/Messaggero Egiziano» del 1° aprile 1937.

Figura 6. Copia scannerizzata formato jpg di «Messaggero Egiziano» del 2 luglio 1929.

4. UN NUOVO INIZIO

Da questo progetto sono scaturite varie iniziative e pubblicazioni[4]: il punto culminante di queste attività, che è stata inaugurata in questi giorni, tratta il periodo alessandrino, in particolare gli anni Trenta, della giornalista, narratrice e attivista politica Fausta Cialente. Esaminando le pagine di «Il Giornale d'Oriente», nel corso dell'elaborazione delle schede descrittive, è emerso il contributo di grande respiro e vario interesse di Cialente, che si può considerare sostanzialmente sconosciuto non solo agli studiosi di letteratura italiana contemporanea, ma anche agli specialisti della scrittrice. Già il primo numero, uscito il primo aprile del 1930, presenta la novella (pubblicata in due puntate) intitolata Guendalina, mia sorella di Fausta Terni Cialente, annunciando la collaborazione nata tra la scrittrice e «Il Giornale»:

«Con questa novella, la signor Fausta Terni-Cialente, notissima scrittrice, vincitrice dell'ultimo Premio dei Dieci, inizia la sua collaborazione al Giornale d'Oriente. Tale preziosa collaborazione continuerà regolarmente poiché a Fausta Terni-Cialente il nostro giornale ha affidato la critica letteraria che verrà iniziata nei prossimi numeri con una cavalleresca recensione della nostra collaboratrice all'ultimo romanzo di Gian Gaspare Napolitano che divise con Natalia l'onore del Premio dei Dieci».

Seguendo il filo di questa produzione giornalistica, ho intrapreso il contatto con il Centro Manoscritti dell'Università di Pavia, dove si conserva un fondo di testi manoscritti e a stampa di e su Cialente, facendo, nello stesso tempo, ricerche sul periodo del suo soggiorno alessandrino. A voler presentare un primo resoconto, è possibile rilevare che le indicazioni autografe dell'Archivio partono dal 1942, le monografie e le ricerche che trattano il soggiorno e le attività cialentiane in Egitto ignorano per lo più gli anni di collaborazione con «Il Giornale d'Oriente», la novella *Guendalina* pubblicata da Cialente sul primo numero non risulta presente nella produzione della scrittrice ripubblicata tanti anni dopo in Italia. E mentre il lavoro procedeva sulla digitalizzazione dei periodici, si è spalancata una finestra sul mondo cialentiano ricco di spunti e prospettive. Per dieci anni, fino al 1940, Fausta Cialente firma, sia per sigla [F.T.C.] sia per nome completo [Fausta

[4] Un mio intervento sarà pubblicato fra gli atti dell'Adi del 2023, con il titolo *Città senza letizia. Il paesaggio naturale e urbanistico egiziano negli scritti di Fausta Cialente su «Il Giornale d'Oriente» (1930-1940)*, dove per paesaggio naturale e urbano si intende principalmente Alessandria e l'Alto Egitto, trattati e descritti, tra il 1936 e 1939, negli articoli firmati da Fausta Terni Cialente e pubblicati, sulle pagine del famoso periodico italo egiziano; un altro mio intervento esce prossimamente negli atti dell'Adi del 2022 sotto il titolo di *Ritorno in colonia. La letteratura italiana del Ventennio in Egitto da «L'imparziale» a «Il Giornale d'Oriente»: una ricognizione*, titolo tratto da una novella di Anna Messina, pubblicata su «Il Giornale d'Oriente», ad Alessandria, nel 1936. Un'altra ricerca importante sarà quella del Dott. Tarek Hussein, che sta elaborando la sua tesi di master sui racconti usciti sulle pagine dell'«Imparziale» dal 1892 al 1930. Le statistiche specificano ben 449 racconti, fra cui tanti inediti, di vari autori italiani che vivevano in Egitto, alcuni anche con nomi fittizi. Un'altra iniziativa vede come protagonisti gli scrittori italiani che viaggiarono in Egitto nei primi decenni del secolo scorso. Sulle pagine di «Le Muse», la bimestrale calabrese d'arte e di cultura, sono stati ripubblicati tra il 2022 e il 2023, le interviste rilasciate con Vivanti e Pirandello durante i viaggi compiuti in Egitto.

Terni-Cialente], vari racconti per «Il Giornale d'Oriente», fra cui *Malpasso, Bambini alla finestra, Paolina, La serva fedele,* tutti del 1937, e *Le statue, Passeggiata con Angela, La ballerina, Albertina* del 1938, *La vedova* del 1939.
Alcuni fra questi racconti compaiono nella raccolta Pamela e la bella estate del 1962. Il contributo di Cialente su «Belle arti e bel mondo/ La pagina letteraria» non si limita alle novelle; la scrittrice pubblica pure articoli che propongono una riflessione sulle opere e sulle figure di scrittori classici e contemporanei, come Anton Čechov, Leone Tolstoj, Torquato Tasso, Sibilla Aleramo, Luigi Pirandello, Gabriel D'Annunzio ed altri. Nell'ambito degli articoli riferibili a fatti di attualità e di cronaca, l'attività di Cialente si svolge sia come redattrice che come inviata speciale in Italia e all'estero. Altri articoli escono sulla pagina di «Cronaca su Alessandria», distinguendo così la versione alessandrina di «Il Giornale d'Oriente» (che usciva ad Alessandria con il sottotitolo di «Messaggero Egiziano»). Su questa pagina, tra il 1936 e il 1939, sono usciti articoli che descrivono e criticano il paesaggio naturale e urbano alessandrino, come: Architettura irrazionale del 3 aprile 1937, e Città senza letizia del 3 novembre 1937; insieme al panorama artistico della città, in particolare riferimento alle mostre dell'Atelier (L'esposizione dell'Atelier del 30 novembre del 1939).
Sono stati elencati, in un anno di ricerche ed esplorazioni, più di cento contributi vari della Cialente, tutti della collezione di «Il Giornale d'Oriente» (1930 – 1940), classificati e ordinati cronologicamente a seconda del genere e del tempo di pubblicazione. È nata, dunque, l'idea di mettere sotto nuova luce questi undici anni che precedono quelli del Diario di guerra e di Radio Cairo, presentarne una nuova e inedita prospettiva in una pubblicazione in cui collaborerò, come ricercatore principale, insieme a Nunzio Ruggiero, Emmanuela Carbé e Laura Cannavacciuolo. Il volume è destinato alla pubblicazione entro quest'anno.

BIBLIOGRAFIA

[1] El Beih, Wafaa, e Emad Saleh. «Sulla documentazione e digitalizzazione del patrimonio culturale italiano pubblicato in Egitto tra il XIX e il XX secolo». *Rivista degli Studi Umanistici e Letterari*, gennaio 2023, 84–103.

[2] Rizzitano, Umberto. «Un secolo di giornalismo italiano in Egitto (1845- 1945)». *Cahiers d'histoire égyptienne* VIII, fasc. 2/3 (avril 1956): 129–54.

Piattaforme wiki per l'insegnamento umanistico: sperimentazioni in corso nel liceo De Cosmi di Palermo

Antonino Fiorino[1], Paolo Monella[2], Francesca Saieva[3], Antonella Sorci[4]
[1]Liceo "G. A. De Cosmi" di Palermo, Italia - fiorino.nino@gmail.com
[2]Sapienza Università di Roma, Italia - paolo.monella@uniroma1.it
[3]Liceo "G. A. De Cosmi" di Palermo, Italia - francescasaieva@gmail.com
[4]AICC, Delegazione di Palermo, Italia - sorciantonella1@gmail.com

ABSTRACT[1]

Negli ultimi anni l'AIUCD si è aperta, in collaborazione con altre associazioni, tra cui l'AICC (Associazione Italiana di Cultura Classica), alla formazione dei docenti di discipline umanistiche nelle scuole, con l'obiettivo di diffondere la cultura digitale e la sperimentazione di metodi didattici digitali, col presupposto che quest'ultima sia più feconda se condotta nella realtà delle scuole. Il presente contributo illustra e riflette su due sperimentazioni in corso presso il Liceo "G. A. De Cosmi" di Palermo, nate anche dallo stimolo di un corso di aggiornamento docenti sull'uso didattico di Wikipedia e Wikisource tenuto da Paolo Monella, organizzato da AIUCD e AICC. Una prima esperienza è stata guidata, negli ultimi due anni scolastici, da Antonino Fiorino: gli studenti vengono guidati a creare o editare pagine Wikipedia riguardanti la letteratura italiana e la storia moderna. Una seconda, guidata quest'anno dallo stesso Antonino Fiorino e da Francesca Saieva, li accompagna nella creazione di un "Atlante del lessico culturale europeo", ovvero un ipertesto, sotto forma di libro nella Wikibooks italiana, con approfondimenti su parole chiave importanti per lo studio della storia, della filosofia e delle letterature europee in chiave interdisciplinare, come "responsabilità, amore, tempo, creatività, divenire" ed altre. Oltre a illustrare i presupposti pedagogici, la rete concettuale e gli obiettivi di queste sperimentazioni, l'intervento ne traccerà un bilancio, sottolineandone i punti critici, gli ambiti di miglioramento e le effettive ricadute sulla formazione degli studenti. In particolare, la creazione di materiali didattici sotto forma di articolo Wikipedia o di libro Wikibook ha un'utilità particolare in un contesto come quello del Liceo De Cosmi. La scuola è infatti collocata in un'area periferica di Palermo, città del Mediterraneo con notevoli problemi di sperequazioni sociali, coesione urbanistica e sociale. Molti studenti del De Cosmi provengono da contesti socioculturali disagiati e non acquistano i libri di testo. Ciò aumenta il valore di ogni sperimentazione didattica che superi l'assetto-classe tradizionale tramite la creazione attiva e collaborativa, anche tramite il digitale, di materiali per l'apprendimento.

PAROLE CHIAVE

Didattica digitale; scuola; Wikipedia; Wikibooks; Liceo De Cosmi di Palermo.

1. INTRODUZIONE

La presente comunicazione presenta due esperienze didattiche basate sull'uso di piattaforme wiki, realizzate al Liceo delle Scienze umane G. A. De Cosmi di Palermo:

1. nella prima, svoltasi nell' a.s. 2022/23, gli studenti (guidati dal docente Antonino Fiorino) hanno creato ed editato pagine Wikipedia su temi di letteratura italiana e grammatica latina;
2. nella seconda, in fase di svolgimento nell'a.s. 2023/24, gli studenti (guidati dai docenti A. Fiorino e Francesca Saieva) elaborano, sotto forma di libro nella Wikibooks italiana, un "Atlante del lessico culturale europeo", con approfondimenti su parole chiave importanti per lo studio della storia, della filosofia e delle letterature europee in chiave interdisciplinare.

La sperimentazione delle piattaforme wiki per la didattica è ormai ben avviata in ambito internazionale: all'interno della assai ampia bibliografia al riguardo, si vedano ad esempio [1, 3, 4, 7, 10, 14, 15]. Per l'Italia si vedano [12, 13], il cap. 11 di [9] e soprattutto [2]. Perdura tuttavia, e non solo nel nostro paese, un pregiudizio in particolare nei confronti di Wikipedia, considerata come sorgente di un facile copia/incolla da parte degli studenti, e per di più come fonte inaffidabile di informazioni (vd. [15]), nonostante la qualità di Wikipedia, o almeno della sua versione inglese, sia stata dimostrata da tempo da uno studio famoso, e anzi ormai persino un po' datato [5].

L'interesse specifico delle due nostre sperimentazioni didattiche risiede in particolare in due aspetti, che saranno messi in mostra nel presente contributo:

[1] Nonostante tutti i co-autori abbiano concordato i principi e l'impostazione del presente contributo, nello specifico i paragrafi 1, 6 e 7 sono stati stesi da P. Monella; il par. 2 da P. Monella e A. Sorci; il par. 3 da A. Fiorino; i par. 4-5 da A. Fiorino e F. Saieva.

1. in primo luogo, essa è nata in seguito ad un corso di formazione docenti presso il Liceo De Cosmi nel 2023, co-organizzato da AIUCD e dalla delegazione di Palermo dell'AICC (Associazione Italiana di Cultura Classica). Essa dunque rappresenta una ricaduta effettiva delle strategie convergenti delle due associazioni: di AIUCD, nella direzione di un allargamento del proprio ambito di attività dalla ricerca alla scuola, soprattutto tramite la formazione docenti; e di AICC, nella direzione dell'apertura al digitale, ambito in cui molti docenti di discipline classiche registrano un ritardo formativo preoccupante;
2. in secondo luogo, la sperimentazione offre un'occasione preziosa per valutare le sfide, le difficoltà specifiche, le ricadute effettive dei metodi didattici sopra richiamati, in un contesto didattico peculiare come quello del Liceo De Cosmi. Questo infatti sorge in un'area periferica di Palermo, città mediterranea con notevoli problemi di sperequazioni sociali, coesione urbanistica e sociale. Molti studenti del De Cosmi provengono infatti da contesti socioculturali disagiati e non acquistano i libri di testo[2]. Ciò aumenta il valore di ogni sperimentazione didattica che superi l'assetto-classe tradizionale tramite la creazione attiva e collaborativa, anche tramite il digitale, di materiali didattici.

2. L'ATTIVITÀ DI FORMAZIONE DOCENTI DI AIUCD E AICC

A partire dal 2021, il direttivo AIUCD ha costituito al suo interno un gruppo di lavoro dedicato alla scuola. L'idea di fondo, nata anche da uno stimolo del presidente emerito Dino Buzzetti, è che la formazione dei docenti di ambito umanistico a metodi didattici digitali rientri tra gli ambiti di interesse delle DH. Sono partiti dunque vari corsi di formazione docenti gestiti direttamente da AIUCD, in collaborazione con altre associazioni accreditate presso il Ministero dell'Istruzione per la formazione docenti, come AICA (Associazione Italiana per l'Informatica e il Calcolo Automatico)[3] e AICC (Associazione Italiana di Cultura Classica), Delegazione di Palermo[4].

L'Associazione Italiana di Cultura Classica opera, attraverso le molte delegazioni in tutta Italia, in piena collaborazione con il mondo della scuola, con quello dell'Università, con istituzioni culturali del territorio, attraverso attività di aggiornamento docenti e iniziative per la valorizzazione e l'attualizzazione del classico.

Con la convinzione che l'alto potenziale di 'comunicabilità' dei testi classici costituisca il punto di partenza per una loro rilettura e una loro valorizzazione nel presente, l'AICC di Palermo ha inaugurato l'anno scorso un progetto 'dal basso' di educazione alle DH nella scuola, i cui punti di forza sono: condivisione con e tra i docenti, fattibilità, realizzazione di pochi e realistici obiettivi, realizzazione di un 'prodotto' emergente da dinamiche metodologiche e didattiche inserite in un effettivo curricolo, ricaduta, replicabilità in contesti differenti; con la finalità (o ambizione) di 'indicare' un modello di apprendimento creativo, organizzativo, democratico, basato sulla collaborazione e sull'interazione tra studenti e docenti.

Dalla collaborazione tra AIUCD e AICC Palermo è nato così un seminario di formazione, tenutosi il 14 aprile 2023 presso il liceo De Cosmi di Palermo. Nel seminario, tenuto da Paolo Monella, coordinatore della commissione scuola del direttivo AIUCD e membro della delegazione AICC di Palermo, dopo un'introduzione generale sui metodi didattici digitali in ambito umanistico[5], è stato illustrato l'uso didattico di Wikipedia e in particolare di Wikibooks. Per quest'ultima piattaforma, l'attività didattica proposta consisteva nel portare gli studenti a creare e modificare edizioni commentate (ed eventualmente tradotte) di testi letterari, come ad esempio quella dei carmi di Catullo[6].

3. CONTESTO DIDATTICO E OBIETTIVI

Il Liceo G. A. De Cosmi di Palermo si compone di diversi indirizzi scolastici: accanto al tradizionale Liceo delle Scienze Umane (indirizzo nativo dell'Istituto) si affiancano il Liceo Linguistico e l'Economico-Sociale. Il contesto socio-culturale dei tre indirizzi, pertanto, si presenta nel complesso vario e multiforme: i titoli di studio dei genitori degli studenti si collocano tra la licenza media e il diploma; mentre sensibilmente diversa è la condizione socio-economica delle famiglie degli studenti iscritti al Liceo Linguistico, mediamente più alta[7]. Inoltre, la zona della città in cui si colloca la scuola non

[2] Per il contesto socio-culturale degli studenti del Liceo De Cosmi, si consulti il Piano Triennale dell'Offerta Formativa (PTOF), https://www.liceodecosmi.edu.it/attachments/article/519/PTOF%20A.S.%202023_2024.pdf, pp. 1-2. Si confronti anche il Rapporto di autovalutazione del Liceo, https://cercalatuascuola.istruzione.it/cercalatuascuola/istituti/PAPM02000N/de-cosmi/valutazione/documenti/, pp. 2-3 e 24-25. Sul potenziale delle tecnologie wiki per 'democratizzare' l'accesso all'istruzione e alle competenze di scrittura di livello alto, si veda ad es. [10].

[3] https://www.aicanet.it. AIUCD ha organizzato due corsi di formazione docenti congiunti con AICA, negli anni scolastici 2022/23 e 2023/24. La liaison tra le due associazioni era costituita proprio da Dino Buzzetti.

[4] https://www.aicc-nazionale.com/ e http://www.paolomonella.it/aiccpalermo/.

[5] Vd. [8] e, più in generale, [9].

[6] Vd. https://it.wikibooks.org/wiki/Carmina_(Catullo).

[7] Vd. il Piano Triennale dell'Offerta Formativa (PTOF) del Liceo, consultabile all'indirizzo https://www.liceodecosmi.edu.it/attachments/article/519/PTOF%20A.S.%202023_2024.pdf, pp. 2 ss.

offre particolari possibilità sul piano culturale, essendo il territorio limitrofo quasi del tutto sprovvisto di centri di ritrovo socio-culturale.
A seguito della pandemia da Covid-19 e della conseguente didattica a distanza attivata a scuola si è riscontrato un sensibile calo del rendimento scolastico degli alunni, ravvisabile in tutte le discipline e in particolare in quelle linguistico-matematiche. Fragilità evidenti, soprattutto nel primo biennio del Liceo che i dati elaborati per il Rapporto di autovalutazione confermano, soprattutto in relazione alle prove standardizzate (INVALSI) che delineano un quadro di fragilità linguistiche e logico-matematiche in linea con i dati del Paese[8]. A partire dal 2022 sono stati attivati strumenti di ausilio e supporto didattico per il recupero delle carenze, miranti a migliorare gli esiti dei livelli di competenza intermedi in uscita e a ridurre gli insuccessi[9].
Negli ultimi anni, sempre con maggiore frequenza e per le più differenti ragioni, non ultime quelle di ordine economico, avviene che gli alunni non siano provvisti del materiale didattico indicato dai docenti (manuali, libri di testo e altro); con la conseguenza che l'azione didattico-pedagogica promossa dai docenti possa risultare poco efficace[10]. Oltre a ciò gli alunni hanno manifestato anche difficoltà di natura metodologica, a cui si accompagna un approccio poco motivato alle discipline. Gli obiettivi che la scuola intende perseguire, pertanto, sono di far maturare negli allievi un approccio orizzontale ai saperi disciplinari 'dal basso', partendo da contenuti essenziali (come ad es. il lessico disciplinare specifico) per giungere a una comprensione globale della disciplina oggetto di studio, anche in un'ottica interdisciplinare.

4. COMPETENZE, METODOLOGIE E DESCRIZIONE DELLE ATTIVITÀ DIDATTICHE

L'analisi dei bisogni educativi e formativi, da cui sono emerse scarsa automotivazione allo studio e fragilità degli studenti nelle discipline linguistico-scientifiche, ha reso evidente la necessità di intervenire per stimolare un apprendimento più dinamico e interattivo da parte degli alunni, assai frequentemente considerati come dei 'vasi' da riempire con i contenuti disciplinari.
Pertanto, nell'a.s. 2022-23, come azione didattica propedeutica allo sviluppo concettuale dell''Atlante del lessico culturale europeo', gli studenti del biennio, cimentandosi nello studio della lingua latina, hanno potuto sperimentare in prima persona le ricadute e i benefici metodologici di un corretto utilizzo di Wikipedia, una piattaforma utilizzata non solo come serbatoio di conoscenze, ma anche come 'luogo' di condivisione di un sapere editabile.
Le competenze digitali degli studenti, attivate e valutate dal docente di riferimento, sono state così ricondotte al quadro europeo del DigComp 2.2 (dai livelli 'base' fino agli 'intermedi')[11]: azioni come l'interazione con le tecnologie digitali in modo ben definito e la scelta dei mezzi di comunicazione digitali per un determinato contesto; la manipolazione di informazioni, dati e contenuti per consentire una migliore organizzazione e recupero e l'individuazione di modalità per creare e modificare contenuti semplici in formati semplici, hanno reso gli alunni, da semplici fruitori di contenuti digitali, costruttori di conoscenze. In tal modo, modificando semplici voci presenti nella piattaforma Wikipedia, in particolare quelle relative alla prima declinazione della lingua latina, gli alunni hanno modificato e 'corretto' alcuni errori facilmente riconoscibile all'interno del corpo del testo: un'azione semplice e immediata, ma dalle importanti ricadute pedagogiche e metodologiche. Da qui, successivamente, l'esigenza di estendere il campo d'indagine e di ricerca alla filosofia e alla storia.
Per il raggiungimento delle competenze sopra descritte, si è intesa l'idea dell''Atlante del lessico culturale europeo' nell'ottica di una scuola del 'dubbio progettuale', nel riconoscimento di saperi resi efficaci dall'operare tecnico-digitale e dalla consapevolezza della costruzione comune di un sapere condiviso (cfr. la piattaforma Wikipedia). Un atlante in forma di wikibook quale 'elogio' della parola e del suo potere comunicativo per orientarsi a un 'modo' di vedere le cose; un ipertesto ove reticolarità e multiprospettivismo si pongono quali strumenti imprescindibili nella costruzione di un sapere aperto e itinerante; un tentativo di sfida alla quotidiana staticità della singola competenza disciplinare (comunemente intesa), ri-abilitata dall'uso di risorse esterne e contestualizzate (materiale didattico multimediale, situazione di partenza, gruppo di lavoro ecc.) e di risorse interne (caratteristiche particolari del soggetto singolo/gruppo, coerenza delle scelte, curiosità e interesse, conoscenze e abilità pregresse ecc.).

[8] Vd. il Rapporto di autovalutazione, e in particolare le pp. 3-6, rintracciabile al seguente indirizzo: https://cercalatuascuola.istruzione.it/cercalatuascuola/istituti/PAPM02000N/de-cosmi/valutazione/documenti/.
[9] Vd. la sezione del PTOF relativa al Rapporto di autovalutazione (RAV), pp. 19 ss.
[10] Alcuni dati sul quadro generale (non limitato alle scuole superiori) del sistema scolastico nella città di Palermo sono in [6], pp. 35-37. Il rapporto URBES 2013 dell'ISTAT rilevava che "Gli studenti palermitani fanno registrare un gap di competenze, sia alfabetica che numerica, rispetto a quelli del Mezzogiorno e – soprattutto – rispetto alla media degli studenti italiani" (vd. (https://www.istat.it/it/files//2013/06/Urbes_2013_Palermo_V_7.4.pdf, p. 131). Dati più aggiornati, specificamente riferiti al fenomeno della dispersione scolastica, particolarmente accentuato nel capoluogo siciliano, sono nella sezione 'Contro la dispersione scolastica' del Portale Scuola del Comune di Palermo (https://portalescuola.comune.palermo.it/?page_id=6025).
[11] https://repubblicadigitale.gov.it/portale/-/digcomp-2.2-il-quadro-delle-competenze-digitali-per-i-cittadini.

La realizzazione dell'atlante coinvolge gli studenti del triennio del liceo G. A. De Cosmi di Palermo. Letteratura italiana e latina, storia e filosofia 'narrate' attraverso parole-stimolo che creano uno spazio globale e 'pubblico' sulla linea del tempo (passato-presente-futuro), raccontando concetti, poetiche, opere e fatti in chiave metacognitiva secondo l'approccio dell'*imparare a imparare*. Parole come 'amore, anima, divenire e tempo' sono finestre grazie alle quali è possibile aprire mondi di conoscenze non limitati alle singole discipline, ma afferenti ai più eterogenei campi del sapere. Un'attività, quindi, che vede protagonisti, tra vecchi e nuovi contenuti, gli studenti guidati dal docente nella ricerca delle fonti, nella verifica dell'attendibilità di queste e nella percezione di una costante e graduale contaminazione culturale che fa da filo rosso nella trama dei saperi.
Il flusso di lavoro così orientato crea le condizioni per il coinvolgimento delle classi del triennio in attività laboratoriali di gruppo nelle ore curricolari, durante le quali per ogni classe è prevista un'attività propria.
In un primo momento, gli studenti di una classe terza (indirizzo linguistico) producono un glossario sulla storia medioevale (il docente suggerisce parole-chiave '*curtis*, feudo, immunità, pellegrinaggio, monarchia feudale, investitura' ecc. e concetti fondamentali dell'epoca storica di riferimento 'poteri universali, pauperismo, oscurantismo' ecc.). Il glossario è stato pensato come strumento funzionale per facilitare (e non semplificare) lo studio della disciplina, integrando piuttosto i contenuti del manuale e, altresì, per valorizzare e potenziare competenze linguistiche. Il glossario, inoltre, nel suo farsi, può avere un effetto domino esteso per la realizzazione di mappe reticolari (macrotesti e microtesti), da realizzare inizialmente in formato cartaceo e successivamente in digitale (documenti Google, *padlet*, piattaforme didattiche ecc.).
Inoltre, gli studenti lavorano all'analisi di testi esemplificativi della letteratura italiana dal Trecento al Romanticismo, producendo un commento analitico realizzato con le risorse della biblioteca scolastica e con gli strumenti digitali dei corpora testuali per rintracciare occorrenze e similarità intertestuali. Gli autori della letteratura italiana, quali ad esempio Dante, Petrarca, Boccaccio, Ariosto, Parini e Foscolo, diventano risorse non solo letterarie ma che lessicali per affrontare tematiche di ampio respiro, aprendo finestre culturali anche sul mondo contemporaneo. Le ricerche lessicali, inoltre, rese possibili attraverso l'uso di specifici strumenti digital, permettono così agli studenti di maturare abilità di analisi quantitativa delle opere e di giungere a una valutazione nel complesso più rigorosa e scientifica dello stile di un autore[12].
Ad altre classi (scienze umane, linguistico e scienze umane con opzione economico-sociale) si propongono una/due tematiche filosofiche e letterarie. Gli studenti si dividono in gruppi di lavoro e, supportati dal docente, iniziano a cercare le fonti digitali (siti culturali, blog, riviste online, video, immagini, scene film ecc.). In questo caso, sono gli studenti a dover trovare le parole-chiave per sviluppare le suddette tematiche, creando in modo transdisciplinare e interdisciplinare più reticoli concettuali possibili; si misurano con questioni trasversali, approfondiscono contenuti e dibattono sugli stessi. Un'attività pensata per l'acquisizione di un lessico filosofico più specifico, per lo sviluppo del pensiero critico, per valorizzare la crescita sul piano relazionale e potenziare le competenze logiche e digitali.
In ultimo a una classe (scienze umane) viene affidato il compito di realizzare dei podcast multimediali letterari, storici e filosofici, di breve durata, che saranno inseriti in un primo tempo in un apposito repository digitale della scuola. La classe costituirà due macrogruppi e dei microgruppi per rendere più agile il lavoro. Gli studenti, dopo essersi adeguatamente documentati (materiale multimediale didattico vario: riviste online, testi digitali, app didattiche, piattaforma per podcast ecc.), produrranno delle interviste immaginando di viaggiare nel tempo con i protagonisti della storia e i filosofi studiati. Un laboratorio per affinare competenze linguistico-digitali e per l'acquisizione di una maggiore consapevolezza della partecipazione attiva al dialogo formativo interdisciplinare.

5. UN BILANCIO

Tracciare un bilancio di questa attività progettuale significa guardare oltre gli orizzonti limitati di una lezione svolta in classe e aprire alle possibilità di quello che si potrebbe realizzare attraverso un utilizzo 'diverso' delle risorse digitali presenti in rete. L'attività digitale di *editing* di voci enciclopediche e di costruzione dell''Atlante del lessico europeo' consente agli alunni di accedere all'infosfera web in modo intuitivo e radicalmente alternativo alle consuete e abituali modalità di fruizione che essi hanno del mondo digitale. L'effettiva possibilità di conoscere il codice sorgente delle piattaforme *wiki* diventa, in tal modo, la chiave di accesso ad abilità e competenze che solo l'informatica declinata nella sua natura umanistica consente di sviluppare.
Le iniziali difficoltà e le ingenue ritrosie a utilizzare lo strumento enciclopedico che gli alunni conoscono fin troppo bene si sono dissipate nel momento in cui ne hanno intuito le potenzialità nella prospettiva di un *editing* pienamente collaborativo. Punto di forza di ogni classe è la grande eterogeneità degli alunni: passioni, interessi, abilità, conoscenze più o meno specifiche che contribuiscono a rendere ogni alunno indispensabile alla realizzazione di un'idea. L'approccio digitale, pertanto, non fa che alzare il livello di partecipazione dell'intero gruppo classe alla costruzione del sapere comune.

[12] Vd. [11].

Se nell'osmosi di dinamiche culturali integrate (umanistica e tecnologico-scientifica) il rischio è la semplificazione, la sfida sta nel ridefinire la complessità. L'organizzazione dei saperi, problematizzata, favorisce, infatti, la fruizione degli stessi tramite una più immediata individuazione dei concetti-chiave restituendo identità e specificità alle singole discipline senza precludere l'interconnessione con altri aspetti culturali, legati all'oggi. *Work in progress*, quindi, per un apprendimento riflessivo, critico e comparato, oltre ogni tradizionale periodizzazione, per comprendere al meglio l'evoluzione delle civiltà e dei loro 'linguaggi'.

Un approccio metodologico prevalentemente empirico, che fa dell'osservazione del dato il punto di partenza e della rielaborazione teorico-ermeneutica il suo punto di arrivo; un arrivo mai definitivo, piuttosto premessa di un nuovo transito spazio-temporale.

Auspicabile, infine, che le due attività didattiche (editing collaborativo di voci su Wikipedia e costruzione dell''Atlante del lessico culturale europeo' su Wikibooks), oltre a potenziare competenze linguistiche, digitali e logiche, a riconoscere nuovi strumenti di lavoro e a formare all'utilizzo critico e consapevole dei *social network* e dei media, rafforzino l'idea di scuola quale agenzia educativa, per una comunità attiva e responsabile attenta alla valorizzazione del dialogo formativo-interculturale.

6. SVILUPPI ULTERIORI

Per quanto concerne la sperimentazione del liceo De Cosmi, mentre il lavoro su Wikipedia è in qualche modo concluso in sé, e i passi successivi consistono nella condivisione 'orizzontale' delle buone pratiche tra i docenti della scuola, per l'Atlante del lessico culturale europeo' il *work in progress* consiste nel suo continuo arricchimento sotto forma di wikibook. Quanto all'attività di formazione docenti di AIUCD e AICC, essa è proseguita in questo A.S. 2023/24 con un altro corso analogo presso il liceo Santi Savarino di Partinico (PA), da cui si attendono ricadute analoghe nell'attività didattica in classe. L'intenzione di entrambe le associazioni è di farne un'iniziativa periodica e stabile.

7. CONCLUSIONI

L'effettiva ricaduta didattica del corso di formazione AIUCD-AICC ha mostrato dunque che una formazione docenti orientata alla sperimentazione di pratiche concrete, pedagogicamente fondate, in cui il digitale non dia solo una marca di novità ma un effettivo valore aggiunto, può avere un impatto positivo. Le sperimentazioni didattiche avviate presso il liceo De Cosmi hanno smentito il pregiudizio che pratiche di apprendimento fondate su strumenti digitali, e su attività complesse come la modifica di voci enciclopediche Wikipedia o la creazione di un wikibook, non siano adatte a contesti svantaggiati dal punto di vista socio-culturale.

BIBLIOGRAFIA

[1] Bruff, Derek. «Students as Producers: Collaborating toward Deeper Learning». In *Scholarship in the Sandbox: Academic Libraries as Laboratories, Forums, and Archives for Student Work*, a cura di Amy S. Jackson, Cindy Pierard, e Suzanne Michele Schadl. Chicago: Association of College and Research Libraries, 2019.

[2] Catalani, Luigi, (a cura di). *Fare didattica con i progetti Wikimedia. Numero speciale della rivista BRICKS.* Vol. 4. BRICKS, 2017. https://www.rivistabricks.it/2017/12/19/n-4-2017-fare-didattica-con-i-progetti-wikimedia.

[3] Foster-Kaufman, Amanda. «Wikipedia-Based Assignments and Critical Information Literacy: A Case Study». In *Critical Approaches to Credit-Bearing Information Literacy Courses*, a cura di Angela Pashia e Jessica Critten, 271–94. Chicago: Association of College and Research Libraries, 2019.

[4] Fulton, Crystal. «The Use of Collaborative Open-Access Publishing via Wikipedia in University Education to Embed Digital Citizenship Skills». *Netcom. Réseaux, Communication et Territoires* 33, fasc. 1/2 (2019). https://doi.org/10.4000/netcom.3893.

[5] Giles, Jim. «Internet Encyclopaedias Go Head to Head». *Nature* 438, fasc. 7070 (2005): 900–901. https://doi.org/10.1038/438900a.

[6] Greco, Gioacchino. *I Presidi e la Scuola Media: Una ricerca a Palermo*. Laboratorio Sociologico. Milano: Franco Angeli Edizioni, 2018.

[7] Mareca, María Pilar, e Borja Bordel. «The Educative Model Is Changing: Toward a Student Participative Learning Framework 3.0 --- Editing Wikipedia in the Higher Education». *Universal Access in the Information Society* 18, fasc. 3 (2019): 689–701. https://doi.org/10.1007/s10209-019-00687-6.

[8] Monella, Paolo. «Didattica digitale e Wikibooks». In *Corso formazione per docenti di discipline umanistiche*. Palermo: AIUCD - AICC - Liceo G. A. De Cosmi di Palermo, 2023.

[9] Monella, Paolo. *Metodi digitali per l'insegnamento classico e umanistico*. Milano: EduCATT, 2020.

[10] Pratesi, Angela, Wendy Miller, e Elizabeth Sutton. «Democratizing Knowledge: Using Wikipedia for Inclusive Teaching and Research in Four Undergraduate Classes». *Radical Teacher* 14 (2019): 22–33. https://doi.org/10.5195/rt.2019.517.

[11] Stoppelli, Pasquale. «La filologia italiana e il digitale». In *Studi e problemi di critica testuale: 1960-2010. Per i 150 anni della Commissione per i testi di lingua*, a cura di Emilio Pasquini, 87–98. Bologna: Commissione per i testi di lingua, 2012.

[12] Tavosanis, Mirko. «Insegnamento universitario della scrittura 2.0 attraverso Wikipedia». In *Tecnologie e metodi per la didattica del futuro. Atti della 27a DIDAMATICA*, 407–10. Pisa: CNR, 2013.

[13] Tavosanis, Mirko. «Scrivere su Wikipedia dall'università alla scuola». In *Scrivere nella scuola oggi. Obiettivi, metodi, esperienze*, a cura di Massimo Palermo e Eugenio Salvatore, 173–82. ASLI Scuola Associazione per la storia della lingua italiana. Firenze: Franco Cesati Editore, 2019.

[14] Wang, Lixun. «Employing Wikibook Project in a Linguistics Course to Promote Peer Teaching and Learning». *Education and Information Technologies* 21, fasc. 2 (marzo 2016): 453–70. https://doi.org/10.1007/s10639-014-9332-x.

[15] Wannemacher, Klaus, e Frank Schulenburg. «Wikipedia in Academic Studies: Corrupting or Improving the Quality of Teaching and Learning?» In *Looking Toward the Future of Technology-Enhanced Education: Ubiquitous Learning and the Digital Nativ*, a cura di Martin Ebner e Mandy Schiefner, 295–311. Hershey, PA, USA: IGI Global, 2010. https://doi.org/10.4018/978-1-61520-678-0.ch017.

The organization and management of the MAGIC project for ancient manuscripts digitization: connections between Mediterranean cultures

Stefania Conte[1], Andrea Mazzucchi[2], Guido Russo[3], Augusto Tortora[4], Giorgia Tortora[5]
[1]University of Naples "Federico II", Department of Physics "Ettore Pancini", Italy - stefania.conte@unina.it
[2]University of Naples "Federico II", Department of Physics "Ettore Pancini", Italy - guido.russo@unina.it
[3]University of Naples "Federico II", Department of Physics "Ettore Pancini", Italy - augusto.tortora@unina.it
[4]University of Naples "Federico II", Department of Humanities, Italy - andrea.mazzucchi@unina.it
[5]Polytechnic of Milan, Department of Physics, Italy - giorgia.tortora@polimi.it

ABSTRACT
This contribution highlights the objectives of the MAGIC project, i.e. the creation of a Service Center that uses cutting-edge technologies for the conservation and valorization of manuscripts and printed texts. The fundamental aim of the project is the usability of digital resources, through modern technologies that can create value, knowledge, comparisons, discussions, and sharing along the shores of the Mediterranean.

KEYWORDS
Digital humanities; digitization; digital conservation; Dante Alighieri; bibliographic resources.

1. INTRODUCTION
The aim of the MAGIC project is to establish a Service Center for technologies applied to the processing of manuscripts, documents and printed texts, specifically digitization, cataloging and creation of metadata.
To achieve this objective, the project combines the multidisciplinary skills of the largest university in the south, the University of Naples Federico II, in particular the scientific knowledge of the Department of Humanities and the "Ettore Pancini" Department of Physics is involved. Different technologies and different software tools are being used in the MAGIC project, and over the time they will be integrated in a unified structure, thus characterizing all the deliverable as coming out of the new MAGIC approach. As far as possible, existing tools will be used, but there are cases, like e.g. the correction for bleed-through effect, in which new, ad hoc tools are being introduced by the MAGIC group. Some of the technologies and tools are listed hereby:
the Internet of Things for the protection and conservation of cultural heritage;
the Internet of Things for the enhancement and use of bibliographic and documentary sources;
digital philology for the acquisition of philological and ecdotic practices applied to medieval and modern literary texts in the digital field;
Artificial Intelligence algorithms for recognizing different types of manuscript writing;
long-term preservation of digitized images, using the F.I.T.S. (Flexible Image Transport System) format;
Big data applied to information retrieval for an accurate access system to digitized materials.
The Service Center is located at the University for experimentation and development, but the intent is to extend the services throughout the national and European territory [7].

2. DIGITAL SPACE IN THE MAGIC PROJECT
2.1. Digital Philology
The MAGIC project is dedicated, in its initial phase, to research in the field of humanistic disciplines, such as philology, cultural heritage sciences and history. The object of interest is a nucleus of 96 illuminated manuscript codices of Dante Alighieri's Divine Comedy, dating back to between the fourteenth and fifteenth centuries (see Fig. 1). The manuscripts come from different cultural institutions, as they are preserved in Italian and foreign libraries and archives. Among others we highlight:
- 1 manuscript of the Biblioteca Nacional de España,
- 1 manuscript of the Bibliothèque et Archives du Château of Chantilly,
- 1 manuscript of the Bibliothèque de l'Arsenal (Bibliothèque Nationale de France),
- 11 manuscripts of the Bibliothèque Nationale de France,
- 2 manuscripts of the Archivio storico civico and Biblioteca Trivulziana of Milano,

- 16 manuscripts of the Biblioteca nazionale centrale of Firenze,
- 1 manuscript of the Biblioteca Nazionale Centrale of Roma,
- 11 manuscripts of the Biblioteca Apostolica Vaticana of Città del Vaticano,
- 2 manuscripts of the Biblioteca nazionale Vittorio Emanuele III of Napoli.

The digitization activity is not the focus of the project, but the production of a broad base of metadata, both relating to the description of the manuscript object and to all existing studies on the subject, which will make the digitized document the point of introduction into history, philology, and manuscript tradition for a complete, multilingual and multicultural journey.

For the first time, a digitalization of content makes all Dante's codes available to specialized and non-specialist users within an online archive and a codicological and iconographic database. The database can process all the metadata concerning the style and attribution of the thumbnails, the iconography of the thumbnails and the relationship between the images and the code text [5].

Indeed, the goal of the MAGIC project is to guarantee the effective readability of all the information content of the manuscripts, through an image quality control system. Ancient manuscripts can be subject to various types of degradation that compromise their readability. Often the oxidation of inks causes the production of acidic substances, which gradually penetrate between the front and back of the page, making the text difficult to read. This effect is known as the bleed through effect.

A technology, already developed, will identify the most suitable and effective techniques to counteract this phenomenon compatibly with its presence and intensity.

The techniques also make it possible to make readable the entire apparatus of comments and glosses, which illuminated manuscripts are rich in [6].

Figure 1. Madrid, Biblioteca Nacional de España, VITR/23/3 (Copyright: www.dante.unina.it)

2.2. Digital Recognition through Artificial Intelligence

Artificial Intelligence (AI) is mentioned several times in the MAGIC project as a tool to assist in the analysis of data obtained from digitization work.

For the HTR (Handwritten Text Recognition) of manuscripts, we compare ourselves with the research underway at the Transkribus project, which presents around 86 public AI models, available for learning the software developed by the project. In this way the best performing orientation can be determined. Given the progress of the development of Transkribus, MAGIC will take advantage of its results, verifying those that already fall within the TRANSKRIBUS models due to formal, typological and historical characteristics and how many others instead require a specific AI model. Automated handwritten character recognition is an open research topic and subject to continuous advances in extraction methods and classification algorithms.

The idea is that an AI algorithm may be able to grasp hidden analogies between apparently different works, on a stylistic level, and this by distinguishing between the set of linguistic facts that make up the style of an author and the set of stylistic

facts which diversify the various phases of the history of a language. This can be done by using various layers of cascaded nonlinear units to perform feature extraction and transformation tasks. Algorithms can be either supervised or unsupervised, and applications include pattern analysis (unsupervised learning) and classification (supervised learning).
AI can be applied directly to the images that make up the manuscripts (e.g. e-Scriptorium), and this makes the AI, and specifically the deep learning algorithms, directly usable on the images of the MAGIC archive, having as input data the various manuscript pages of the same author or images of manuscripts by different authors, but on similar topics.
An application of AI is in the removal of the bleed-through effect. Whilst tools like Transkribus and e-Scriptorium can perform digital recognition of handwritten documents using machine learning techniques, they are not usable to simply "clean" the images of the bleed-through effect, an action desired by those who want to maintain the image without digital recognition of characters, just to make the image more "readable" (increase fruibility). This is done with a new tool developed within the MAGIC collaboration, thus implementing an integrated approach in which several tools are used, depending on the user request (see Fig. 2) [2, 4].

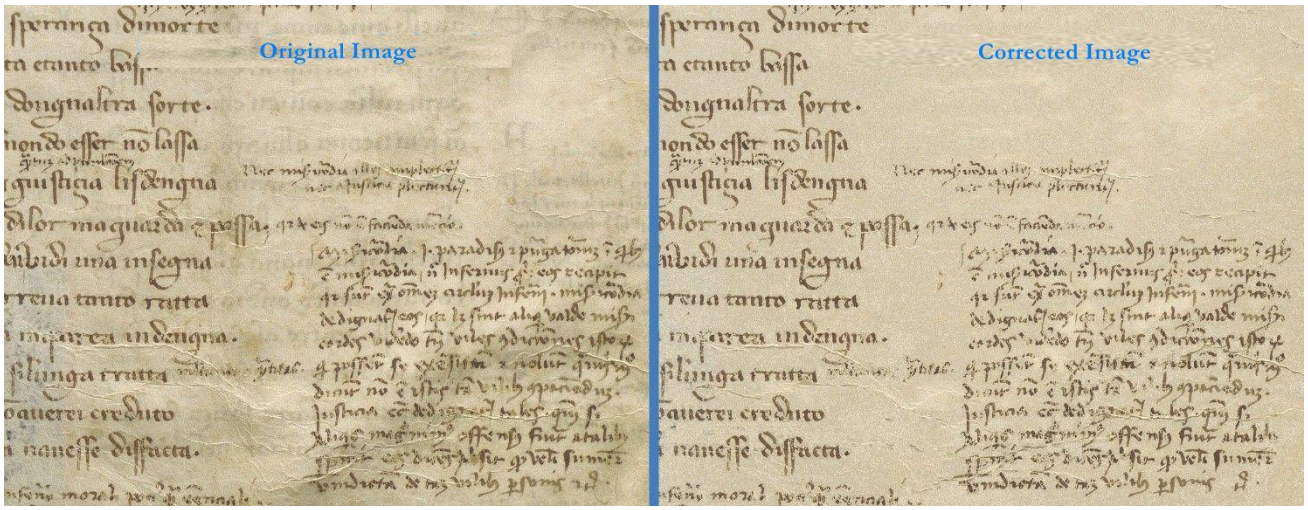
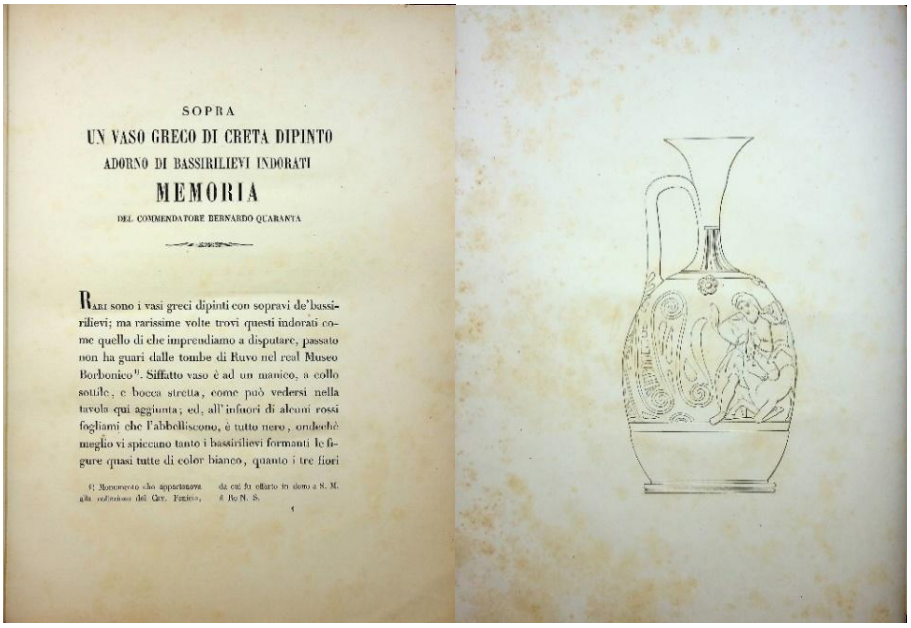


Figure 2. Reduction of the bleed-through effect

2.3. Digital Collection

Interest of the Magic project is a further digitization activity. The aim is to create a virtual collection of the 15th and 16th century editions which are preserved in the library of the Accademia Pontaniana, a Neapolitan academy, founded around 1443. The object of the digitization is the digital acquisition of 6 incunabula and 186 16th century works, coming from the book collection belonging to Francesco and Luigi Torraca. A book collection that creates convergences and interactions between Mediterranean peoples, civilizations and cultures: the collection brings together works of Greek, Latin, Jewish and Italian literature, as well as texts on history, law, philosophy, religion, law, architecture, medicine, numismatics, astronomy and geography.
During the digitization prototyping phase, a printed monograph of the Academy entitled "Memorie della regale Accademia ercolanese di archeologia" was digitized (see Fig. 3). The monograph, dating back to 1862, collects the classical studies, accompanied by illustrations that were conducted by members of the Academy, regarding the artistic influences in the Neapolitan civilization and, therefore, the proximity to Greek culture.
This will be followed by the digitization of selected incunabula belonging to one of the richest libraries in Naples, the Girolamini Library which includes various book collections, from which works of literature, philosophy, Christian theology, philosophy, history of the Church and sacred music come. The agreement with the library is being defined for this activity.
The digitization of the texts of the Pontaniana Academy and the Girolamini Library is carried out at high resolution using planetary scanners and according to the guidelines of the International Federation of Library Associations and Institutions (“Guidelines for Planning the Digitization of Rare Book and Manuscript Collections, 2015”).



SOPRA

UN VASO GRECO DI CRETA DIPINTO

ADORNO DI BASSIRILIEVI INDORATI

MEMORIA

DEL COMMENDATORE BERNARDO QUARANTA

RARI sono i vasi greci dipinti con sopravi de' bassirilievi; ma rarissime volte trovi questi indorati come quello di che imprendiamo a disputare, passato non ha guari dalle tombe di Ruvo nel real Museo Borbonico[1]. Siffatto vaso è ad un manico, a collo sottile, e bocca stretta, come può vedersi nella tavola qui aggiunta; ed, all'infuori di alcuni rossi fogliami che l'abbelliscono, è tutto nero, ondechè meglio vi spiccano tanto i bassirilievi formanti le figure quasi tutte di color bianco, quanto i tre fiori

1. Monumento che appartenne alla collezione del Cav. Fonton, da cui fu offerto in dono a S. M. il Re N. S.

Figure 3. Printed volume "Memorie della regale Accademia ercolanese di archeologia" (Copyright Accademia Pontaniana)

After scanning, a directory is produced for each individual volume containing images in TIFF format in very high resolution corresponding to the individual pages of the work. The TIFF format is the most suitable solution due to its extreme flexibility and the possibility of using lossless compression techniques to reduce files and make data use more effective. In fact, for each of the scans the JPG format is also produced, suitable for consultation in person and online [3].

2.4. Digital Preservation

As described above, images are acquired by the scanner in TIFF and JPG formats. Furthermore, via the back-end interface it is possible to transform the same images into F.I.T.S. format. (Flexible Image Transport System). Even in the prototype phase, digital acquisition in TIFF and JPG formats was followed by their conversion into F.I.T.S. format. The F.I.T.S. format, designed by NASA in the 1970s and used by the Vatican Apostolic Library, is an open standard that has the objective of long-term archiving and relative transmission of images, in the name of the principle "once F.I.T.S., always F.I.T.S." This means that the data saved in this format will always be consultable and compatible with any evolutions of the standards.

The structure of a F.I.T.S. file is very simple, but effective. A file is composed of two distinct parts, which can be repeated several times in sequence:

"Header", in ASCII characters, contains the description and information on the data, in binary form, i.e. the keywords (author, format, dimensions, history of the book object, observations, etc.);

"Data Unit".

Essentially, the file is self-documented because the metadata is included within it.

A back-end interface will be created via a private server and with a minimal interaction method, which provides the operator with an immediate method to:

1. insert the scanned images in TIFF and JPG;
2. transform the TIFF and JPEG files into the F.I.T.S. format;
3. carry out the data entry keywords operation in the header of the F.I.T.S.;
4. upload images to the storage system.

In 2022, the Italian National Unification Body recognized the F.I.T.S. as a standard format for the digital preservation of ancient books and manuscripts (UNI 11845, Processes for managing long term preservation of digital images using the FITS file format).

Concerning the conservation aspects, the project started characterizing the materials (micro particles of paper, ink) by means of an FT-IR Spectrometer, aiming at determining the long-term problems of the volume. This means that MAGIC will give a contribution to book conservation, but not actually do it, as this is reserved to the book owner [1].

3. ACKNOWLEDGEMENTS

The project was funded by Ministry of Enterprise and Made in Italy, (code n. F/130093/03/X38 and CUP: B69J23000560005) and by Ministry of University and Research (code n. PIR01_00011, CUP: I66C18000100006).

REFERENCES

[1] Allegrezza, Stefano. «Analisi del formato FITS per la conservazione a lungo termine dei manoscritti. Il caso significativo del progetto della Biblioteca Apostolica Vaticana». *DigItalia* 6, fasc. 2 (2011): 43–72.

[2] Conte, Stefania, Gian Marco Di Domenico, Andrea Mazzei, Andrea Mazzucchi, Guido Russo, Alessandro Salvi, e Augusto Tortora. «The MAGIC project: first research results». In *Proceedings of the 20th Conference on Information and Research science Connecting to Digital and Library science. Bressanone, Brixen, 22-23 February 2024*, a cura di Eleonora Bernasconi, Andrea Mannocci, Antonella Poggi, Angelo Salatino, e Gianmaria Silvello, 87–93, 2024.

[3] Conte, Stefania, Pasqualino M. Maddalena, Andrea Mazzucchi, Leonardo Merola, Guido Russo, e Guido Trombetti. «The role of project MA.G.I.C. in the context of the European strategies for the digitization of the library and archival heritage». In *Eurographics Workshop on Graphics and Cultural Heritage*, a cura di Alberto Bucciero, Bruno Fanini, Holger Graf, Sofia Pescarin, e Selma Rizvic, Eurographics Workshop on Graphics and Cultural Heritage:119–28. The Eurographics Association, 2023. https://doi.org/10.2312/gch.20231167.

[4] Feliciati, Pierluigi. «Archival users and AI tools for reference and access: a study within the InterPARES Trust AI project». *JLIS.It* 14, fasc. 3 (2023): 117–28.

[5] Mazzucchi, Andrea. «Alterità, leggibilità e traducibilità nella letteratura italiana medievale. Se siano sufficienti i "contenuti di realtà" per recuperare la fruibilità dei testi medievali». *Medioevo romanzo* XL, fasc. 1 (2016): 169–83.

[6] Mazzucchi, Andrea, e Enrico Malato. *Censimento dei commenti danteschi. 1. I commenti di tradizione manoscritta (fino al 1480)*. Roma: Salerno Editrice, 2011.

[7] Russo, Guido, Luciano Aiosa, Giancarlo Alfano, Angelo Chianese, Fabio Cornevilli, Gian Marco Di Domenico, Maddalena M. Pasqualino, et al. «MA.G.I.C.: Manuscripts of Girolamini in Cloud». *IOP Conference Series, Materials Science and Engineering,* 949, fasc. 012081 (2020): 1–8.

Towards a resemantisation of the concept of modelling in Digital Humanities

Cristina Marras[1], Arianna Ciula[2], Øyvind Eide[3], Patrick Sahle[4]
[1]CNR-ILIESI, Italy - cristina.marras@cnr.it
[2]King's Digital Lab, Great Britain - arianna.ciula@kcl.ac.uk
[3]Universität zu Köln, Germany - oeide@uni-koeln.de
[4]Bergische Universität Wuppertal, Germany - sahle@uni-wuppertal.de

ABSTRACT

The presentation focuses on a "conceptual map" of selected key concepts made by the authors to underpin the theory and practice of modelling in the Digital Humanities (DH). The map, associated glossary and examples of practice are derived from the authors' collaborative research and the recently published volume: *Modelling between digital and humanities: thinking in practice*. Departing from the book, we aim to dwell further on a few points. We would like to propose a resemantisation of the concept of model, guided by the idea that a language on modelling in DH can be developed by mapping relevant uses of the term so as to grasp not only the theoretical but also the practical dimension. Terminological reflections help us to highlight the dynamics and interdependence between the two dimensions and, above all, their interdependence, but can also foreground the challenges of connecting any theoretical work back to the actuality of models. Background assumptions for the paper are the reliance on the mediating capacity of language in the construction of interdisciplinary spaces and the central role of modelling in DH as a pragmatic process of thinking and reasoning where meaning is negotiated through the creation and manipulation of external representations combined with an imaginative use of formal and informal languages.

KEYWORDS

Terminology; conceptual map; reasoning strategy; interdisciplinarity.

1. INTRODUCTION

This paper focuses on a "conceptual map" of selected key concepts made by the authors to underpin the theory and practice of modelling in the Digital Humanities (DH). The map, associated glossary and examples of practice are derived from the authors' collaborative research and the recently published volume: *Modelling between digital and humanities: thinking in practice* [2][1]. Background assumptions for the paper are the reliance on the mediating capacity of language in the construction of interdisciplinary spaces and the central role of modelling in DH as a pragmatic process of thinking and reasoning where meaning is negotiated through the creation and manipulation of external representations combined with an imaginative use of formal and informal languages.

Departing from the book, we would like to elaborate on a few points. First, we would like to propose to colleagues a kind of resemantisation of the semantic field of the concept of "model/s", guided by the idea that a language about modelling in DH can be developed through the mapping of relevant uses of the term [3, 5]. By resemantisation we intend here a process by which words or expressions are transformed to denote a specific concept for specific purposes (terminologisation). As a result of this resemantisation, the set of terms that emerged from our research on models and modelling takes on a new meaning in the specific context of modelling in DH, understood as 'thinking in practice', in order to capture not only the theoretical but also the practical dimension (§2). Secondly, the paper brings to the fore a reflection on terminology to highlight the dynamics and interdependence between the two dimensions and, above all, their interdependence. In this way, the map also highlights modelling as a strategy of reasoning that makes it possible to overcome disciplinary boundaries[2], but also make the challenge of linking any theoretical work back to the actuality of models emerge (§3).

[1] The book also traces the debate about models and modelling in the DH, please refer to the volume for the bibliography [2: 207-223], see also [1, 7 and 9].

[2] There is currently a particularly lively discussion on multilingualism in DH: cfr. https://multilingualdh.org/en/; see also: https://zenodo.org/communities/multilingual-dh/. "Digital multilingualism is not just about linguistics, or language. It has a cultural and socio-political dimension which is crucial in studying increasingly transcultural and translingual dynamics and in helping us to understand what are ultimately human-designed and complex (digital) cultural artefacts" [8].

2. FROM TERMS TO CONCEPTUAL MAP

The conceptual map presented in Fig. 1 is focused on the term "model/s". The map can be analysed and navigated from two different (interrelated) perspectives: focusing on the terms (boxes), and focusing on the relationships (edges) between the selected terms; the latter is an attempt at modelling a modelling process [5].

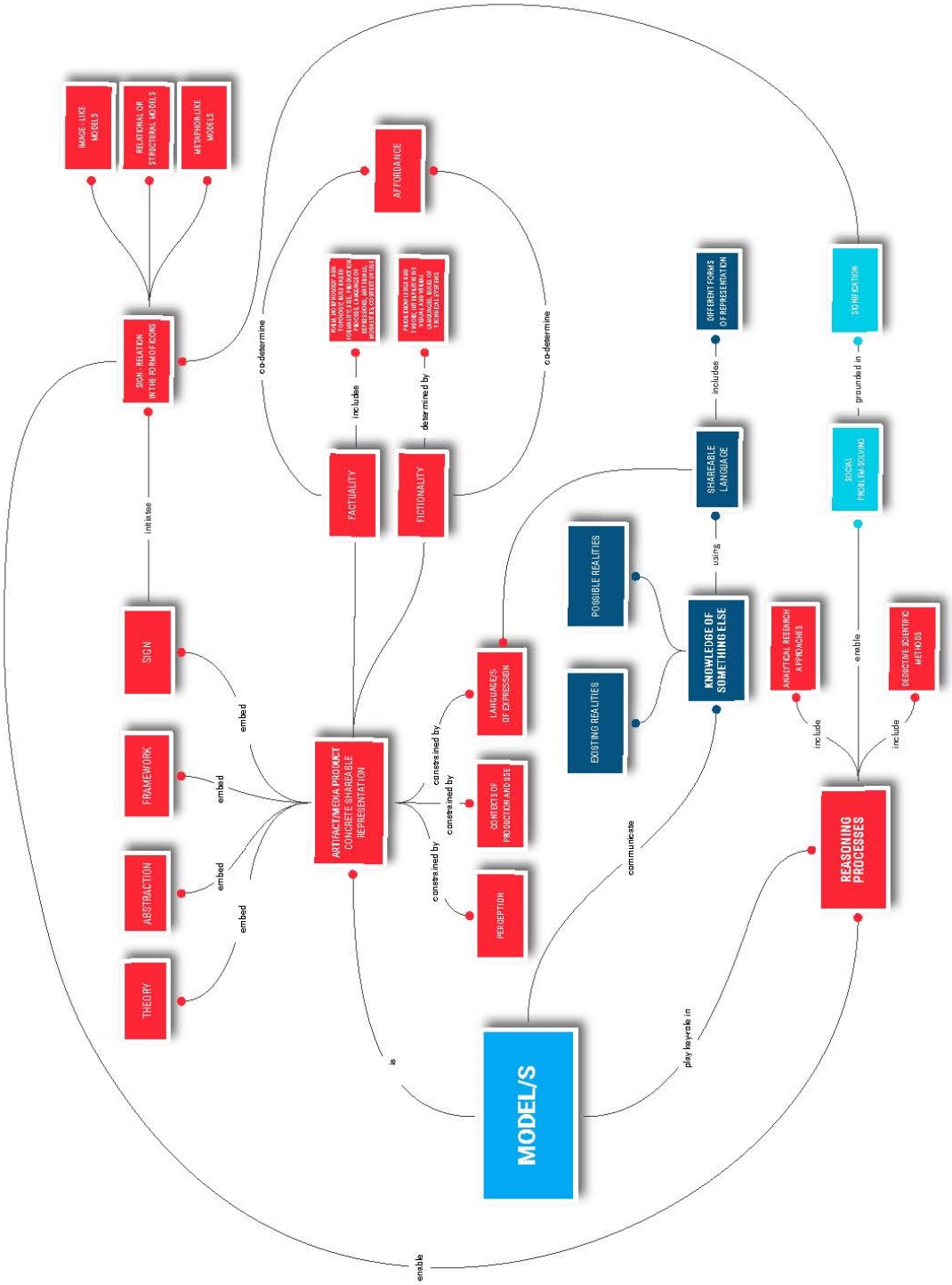


Figure 1. Conceptual map of a selection of the key terms[3]

[3] The map was designed by Silvestro Caligiuri (CNR-ILIESI).

From the term “model/s” a series of elements branch off that, according to the path of our research, build the fabric on which modelling as a process unfolds; these elements are the warp and the relations the weft of this articulated fabric or, recurring to a different metaphor, the different elements are the building blocks that make up the different aspects and components of modelling.[4] Building on the literature and according to this map, “model/s” plays a key role in *reasoning processes*[5] and knowledge development and sharing. “Model/s” is a heuristic tool by means of which an object is re-described as a result of a modelling process; it is at the same time an *artefact*, hence a media product, a concrete (visual, perceptible) shareable representation or expression embedding an element of theory, abstraction, a framework, or a sign. “Model/s”*communicate knowledge of something else* using shareable language including different forms of representation. Models are media products mediated by the conditions and constraints of their perception but also by its language/s of expression; in light of the latter, metaphors assume a central role where meaning is negotiated through the creation and manipulation of external representations combined with an imaginative use of languages with different levels of formalisation and modes of expression.

Models are contingent, created in actual scholarly situations of production and use; partially arbitrary in that the same inferences drawn by manipulating one model could have been reached in other ways, for instance using a different model. Factuality (form, morphology and topology, rule-based formality, context of use) and fictionality (subjectively determined dependency on prior knowledge and theory) of models are entangled and concur to determine their affordances.

“Models as concrete forms with an identifiable level of formality provide affordances to the intellectual process by enabling and constraining the development of what can be represented and how it can be represented” [2: 128]. Translation could be considered the 'engine' of the modelling process, in the sense that the process of signification that unfolds in modelling activities implies translation, negotiation and transformation of meaning. We experience this, for example, when we are modelling texts [2, chapter 4 and 5], one of the most common objects of modelling activities in the DH tradition.

This mediation is particularly evident in the tangible physical forms of models (material and mediated media products), in which the act of modelling includes the technical apparatus via which models are operationalised and interpreted [4]. Formal and experimental modelling techniques are often combined with a constructive use of programming and verbal and visual languages (see Fig. 2). Specific and situated visual representations are one of the primary results of this translation process.

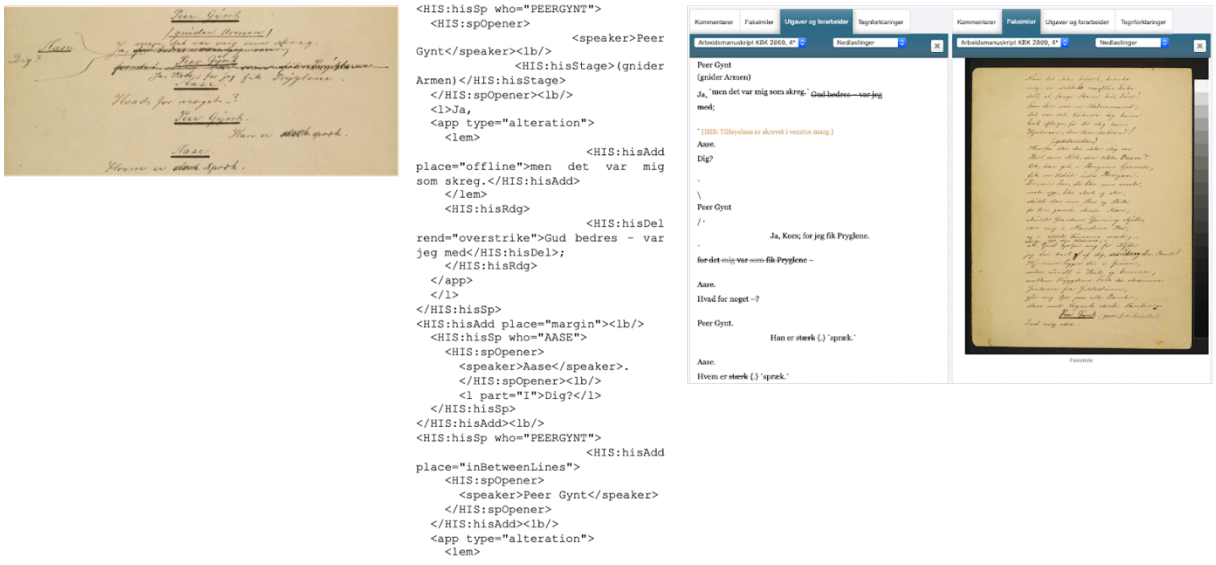


Figure 2. Media transformation between media products: from text to model [2: 116]

An example on how we can model concepts of “text” is shown in Figures 3 and 4. A textual model (a narrative) is translated into something that looks “somewhat” formal (diagrammatic, relational/structural model).

[4] From the very beginning of our research project, we resorted to the metaphor of construction, including the use of legos to 'represent' our work hypothesis and the process of modelling / thinking in practice [2: 2, Fig. 1].

[5] This claim applies to both formal ways of reasoning and representation, pertaining to deductive scientific methods, and to less formal ones, mostly attributable to analogical research approaches.

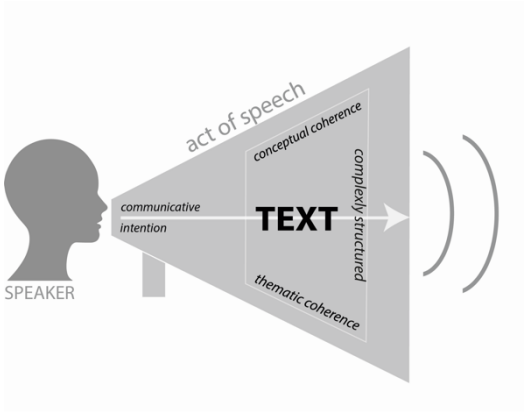


Figure 3. From a short definition of 'text' (bib147) "Ein Text ist eine komplex strukturierte, thematisch wie konzeptionell zusammenhängende sprachliche Einheit, mit der ein Sprecher eine sprachliche Handlung mit erkennbarem kommunikativem Sinn vollzieht". Trad. "A text is a complexly structured, thematically as well as conceptually coherent linguistic unit, with which a speaker executes a verbal action with recognisable communicative sense" [2: 147].

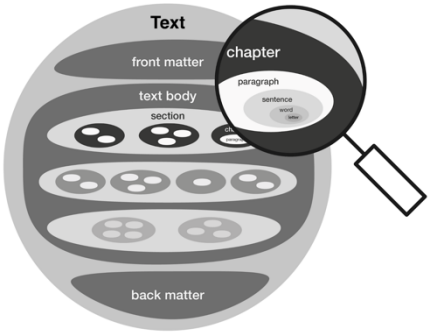


Figure 4. "Text is an ordered hierarchy of content objects". The OHCO model [2: 183].

Back to the map (see Fig. 1): the semantic field that unfolds in the map highlights the processes of conceptualisation, representation, visualisation and communication that characterise the term model, reflecting the polysemy rooted in its etymology. This polysemy also gives rise to the multiplicity of properties that, as the map shows, are progressively articulated around the model and the network of relations associated with it (resemantisation). This focus on language and terminology frames modelling as a process of signification (semiotic process or meaning making). In turn, this semiotic framework allows us to see modelling primarily as a strategy to make sense (signification) via practical thinking (creating and manipulating models). As it emerges from the map, the semiotic perspective can be complemented by modelling studied as a media transformation process. The reinstantiation in renewed terms (terminologisation) of the understanding of modelling is an attempt to respond to the dynamic nature of models and modelling as an open process of signification that enacts a triadic cooperation (between object, representamen and interpreter). This understanding of modelling could ultimately allow us to surpass the rigid duality object vs. model, as well as sign vs. context.

3. MODELLING AS A RESEARCH STRATEGY FOR INTERDISCIPLINARITY

Modelling in DH emerges from its language and metalanguage as a pragmatic process of thinking and reasoning where meaning is negotiated through the creation and manipulation of external representations combined with an imaginative use of formal and informal languages [2, 6]. From this process, modelling emerges as a reasoning strategy capable of overcoming disciplinary boundaries, but also making the challenge of connecting theoretical work back to the actuality of models emerge [8, 9]. In our presentation we propose a sort of resemantisation of the concept of model guided by the idea that a language on modelling in DH can be developed through the mapping of relevant uses to grasp not only the theoretical but also the practical dimension. A challenge for a research agenda in this area would be to explore how the interplay between intrinsic structures of models (selection of salient qualities) and extrinsic mapping (their iconic ground) develops in the creation of scholarly arguments in the humanities.

The conceptual map building on our research on modelling showcases conceptual pathways (semantics) and the elements articulating the modelling process and uses (pragmatics). The dynamics and interdependence between these theoretical and practical dimensions, highlight their mutual influence. Any discussion of modelling – given its porosity in many fields of study – must address the context of interdisciplinarity and the opportunities and weaknesses of polysemy that have developed throughout the history of the use of terms. In fact, model and modelling have been and can be considered objects of study also in DH. Placing this analysis at the intersection of multiple disciplines with many points in common (cross-disciplinarity) facilitates the transfer of methods from one discipline to another (trans-disciplinarity). This approach integrates (inter-disciplinarity), rather than assembling and separating, disciplines, data, methods, tools, theories, in order to create a common understanding.

REFERENCES

[1] Beynon, Meurig, Steve Russ, e Willard McCarty. «Human Computing--Modelling with Meaning». *Literary and Linguistic Computing* 21 (1 giugno 2006). https://doi.org/10.1093/llc/fql015.

[2] Ciula, Arianna, Øyvind Eide, Cristina Marras, e Patrick Sahle. *Modelling between Digital and Humanities. Thinking in Practice*. Open Book Publisher, 2023. https://doi.org/10.11647/obp.0369.

[3] Ciula, Arianna, Øyvind Eide, Cristina Marras, e Patrick Sahle. «Models and Modelling between Digital and Humanities - A Multidisciplinary Perspective». *Historical Social Research, Supplement* 31 (2018). https://www.gesis.org/en/hsr/current-issues/2018/suppl-31-models-and-modelling-between-digital-and-humanities.

[4] Elleström, Lars. «The Modalities of Media: A Model for Understanding Intermedial Relations». In *Media Borders, Multimodality and Intermediality*, a cura di Lars Elleström, 11–48. London: Palgrave Macmillan, 2010. https://doi.org/10.1057/9780230275201_2.

[5] McCarthy, Willard. «Modeling: A Study in Words and Meanings». In *A Companion to Digital Humanities*, a cura di Susan Schriebman, Ray Siemens, e John Unsworth. Blackwell, Oxford, 2004, 2004. http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-7.

[6] McCarthy, Willard. «Modelling What There Is: Ontologizing in a Multidimensional World». *Historical Social Research, Supplement* 31 (2018): 33–45. https://doi.org/10.12759/hsr.suppl.31.2018.33-45.

[7] Orlandi, Tito. *Informatica umanistica*. Roma: Carocci, 1999.

[8] Spence, Paul. «Disrupting Digital Monolingualism: A report on multilingualism in digital theory and practice». Language Acts & Worldmaking project, 2021. https://doi.org/10.5281/zenodo.5743283.

[9] Thaller, Manfred. «From History to Applied Computer Science in the Humanities». *Historical Social Research, Historische Sozialforschung, HSR Supplement* 29 (2017). https://www.gesis.org/en/hsr/full-text-archive/2017/suppl-29-from-history-to-applied-computer-science-in-the-humanities.

Valorizzare un archivio ‘mediterraneo’: studi per un’edizione critica digitale delle opere di Giovanni Comisso

Marco Borrelli
Università degli Studi di Napoli “L’Orientale”, Italia - marco.borrelli@unior.it

ABSTRACT

All’interno di un più ampio percorso di valorizzazione e digitalizzazione dei depositi memoriali di area mediterranea del XX secolo, promosso dal centro di studi ALMA, s’intende dare la giusta visibilità alle collaborazioni giornalistiche e alla produzione letteraria di un autore rimasto un po’ ai margini del canone letterario del Novecento, quale Giovanni Comisso. In prima istanza, ripercorrendo le tappe della sua formazione e gli episodi significativi che influenzano la sua prima stagione narrativa, si pone l’attenzione sulle varie e valide ragioni per cui questo scrittore può essere assunto come punto di riferimento per la definizione di una moderna storia culturale del Mediterraneo. In secondo luogo, per rendere noto lo *status quaestionis* inerente alla digitalizzazione dei suoi scritti, si offre una panoramica del lavoro svolto dall’Associazione Amici di Giovanni Comisso, diretta da Ennio Bianco, mostrando brevemente le funzionalità del *Centro di Documentazione Digitale Giovanni Comisso*. Infine, in vista di un’ulteriore opera di elaborazione e organizzazione dei materiali presenti nell’archivio dello scrittore, si ipotizza una DSE de *Il porto dell’amore* e si mostra, a titolo esemplificativo, una prima trascrizione in XML/TEI predisposta per l’edizione critica digitale della raccolta *Gente di mare*, in linea con i nuovi paradigmi della filologia digitale.

PAROLE CHIAVE

Comisso; Mediterranean Studies; Storage; Scholarly Digital Edition; Short Stories.

1. COMISSO TRA SCRITTURA MEMORIALISTICA E GIORNALISMO

La lunga e varia attività di Giovanni Comisso, nato a Treviso nel 1895, attraversa gran parte del Novecento fino agli anni del boom economico, passando per numerose collaborazioni a riviste e giornali, tra cui, per citarne alcuni, “Il Convegno”, “Il Mondo” e il “Corriere della sera”, dal quale viene ingaggiato per scrivere un reportage sull’Estremo Oriente, pubblicato poi in una serie di articoli tra il gennaio e l’agosto del 1930 [15: 9-11]. Questi testi nati su commissione costituiscono un caso filologico interessante e restituiscono la cifra del *modus operandi* di Comisso, che a distanza di tempo torna su questi materiali per dar vita a un’opera completamente diversa, scevra dalle imposizioni morali e dalle *contraintes* dei giornali: *Amori d’Oriente*. La carriera del trevigiano raggiunge l’apice della notorietà nel 1955, quando con i racconti di *Un gatto attraversa la strada* si aggiudica la vittoria del Premio Strega. Dopo una giovinezza avventurosa, Comisso sostanzia il sogno di una vita contemplativa acquistando una villa a Zero Branco, località immersa nella campagna veneta, dove all’insegna di antiche tradizioni arcadiche può trascorrere le proprie giornate in sintonia con i cicli della natura. Quest’esperimento di vita bucolica restituisce l’idea di quanto lo scrittore sia fortemente legato alla propria regione, eppure, come lasciano intuire i tanti viaggi in giro per il mondo e la stessa collocazione geografica di Treviso, affacciata sull’Adriatico e sull’Istria, egli non vive in uno spazio di isolamento culturale, piuttosto appare fin dagli esordi un autore ribelle ai ‘confini’ [22: 41]. Rileggendo quasi freudianamente alcuni eventi dell’infanzia, Comisso stesso suggerisce di interpretare la propria vita a partire da questo dualismo. L’aneddoto risale a quando da infante, a causa della scarsità del latte materno, viene affidato a due balie differenti; in questo suo primordiale ‘vagabondaggio’ egli coglie un significato allegorico che illumina *a posteriori* la sua esistenza: «questo mio errare è stato lo schema prestabilito del continuo mio muovermi per tutta la vita da un paese all’altro pure avendo invece il desiderio di stare fermo in incanto e contemplazione» [6: 27]. Il desiderio per l’incanto della contemplazione è quindi il rovescio della medaglia di uno spirito anelante a una libertà selvaggia, che ha trovato sfogo nella partecipazione alla Gande Guerra prima e nell’adesione alla celebre impresa fiumana poi. A Fiume, tra l’altro, conosce la ciurma del capitano Gamba, al bordo del cui bragozzo si imbarcherà più volte tra le estati del 1922 e del 1928, come testimoniato dalle cartoline e dalle fotografie consultabili tra i materiali custoditi presso il Fondo Comisso della Biblioteca Comunale di Treviso. Tutte queste esperienze accentuano in lui una vocazione transnazionale, che gli consente di andare oltre i confini italiani e di guardare agli Stati limitrofi non tanto come presenze antagonistiche, bensì come aree di una macroregione fondata sulla coabitazione di culture, nella quale le diversità sembrano affievolirsi a beneficio di una condivisione degli spazi mediterranei. Una bellissima testimonianza è offerta dal racconto di

una traversata adriatica, intentata insieme all'aviatore Guido Keller a bordo di un'imbarcazione sgangherata, per raggiungere la Morlacchia [7: 102-104].
Per un autore che ha trovato nella scrittura memorialistica la sua vena più autentica, come più volte sottolineato dalla critica avallando alcune affermazioni dell'autore[1], vita e scrittura spesso tendono a sovrapporsi e a confondersi: tant'è che «è impossibile scindere la biografia […] dall'opera» [8: 12]. Infatti, dagli eventi risolutivi della sua giovinezza, Comisso trae ispirazione per la pubblicazione delle prime opere in prosa: *Giorni di guerra*, *Il porto dell'amore* e *Gente di mare*. Queste ultime due, benché basate su esperienze che cronologicamente seguono quella della Grande Guerra, sono edite prima di *Giorni di guerra*. Con ogni probabilità, per riuscire a trovare l'equilibrio formale e stilistico necessario per il racconto dell'esperienza bellica – che per di più è completamente avulsa dalle coordinate della memorialistica fascista, votata all'esaltazione dell'eroismo italiano – Comisso sente l'urgenza di instaurare un rapporto fecondo con il mare, di confrontarsi con il Mediterraneo e trovare delle risposte che lo conducano al di là delle ideologie e del dannunzianesimo.

2. IL RICHIAMO DELL'ADRIATICO: *IL PORTO DELL'AMORE*

Dopo l'esordio letterario avvenuto nel 1916, con un libretto di poesie edito grazie all'interessamento dello scultore Arturo Martini, la prova narrativa con cui Comisso si fa conoscere dal grande pubblico si presenta coma una cronaca lirica dell'occupazione di Fiume, basata sulla rievocazione di quell'«atmosfera di fascinazione che rompe gli schemi della storiografia» e che Carlo Federico Simonelli definisce altresì con l'espressione «poesia in movimento» [21: 5]. *Il porto dell'amore* ripercorre, facendo ricorso a una struttura episodica, alcuni aspetti della vita anticonformistica condotta a Fiume – tra il 12 settembre del 1919 e il Natale del 1920 – dall'autore insieme ad altri giovani legionari, partiti al seguito di D'Annunzio e protagonisti degli eventi culminanti con la proclamazione della Reggenza italiana del Carnaro [9]. Fin da questi racconti, è ben visibile come in Comisso la letteratura non sia mai discosta dalla vita: la penna dello scrittore si rivela sempre pronta a cogliere lo svolgersi dell'esistenza nella sua immediatezza, eliminando ogni diaframma tra le proprie percezioni multisensoriali e la pagina che ha davanti. A tal proposito, in una recensione comparsa sul periodico "Il Quindicinale" nel 1926, Montale, tra i primi ad accorgersi della freschezza dello stile comissiano, scrive che *Il porto dell'amore* è un «libretto carnale e febbrile, che avvampa e trascolora è appena un libro ed è ancora una malattia. Arte legata alle primavere del sangue, al corso delle stagioni e alle temperie: poco più di un rabesco, il diagramma di una vita rovesciata sulle cose; ma d'una purezza, talvolta, di cristallo» [4: 1628]. Insistendo sugli aspetti estatici e visionari del libro, il poeta genovese indica i tratti stilistici che avvicinano l'autore alla cultura francese di fine Ottocento – ai simbolisti e a Rimbaud su tutti – nonché ai *Canti Orfici* di Dino Campana. Con Comisso, in altre parole, si è dinanzi a uno di quei casi in cui l'esistenza avventurosa si riflette nelle scelte artistiche: è proprio il richiamo del mare, emblema della libertà assoluta, a spingerlo a varcare la soglia del dannunzianesimo divenuto ormai imperante nella letteratura italiana. Difatti, già nel periodo fiumano, come sarà poi ben visibile tanto in alcuni testi presenti ne *Il porto dell'amore* quanto nelle pagine a questo periodo dedicato nella successiva autobiografia *Le mie stagioni*, Comisso prende le distanze dal vate cercando una strada filosofico-artistica alternativa, che trova espressione nella fondazione della rivista fiumana "Yoga". La vita da legionario spesso lo annoia e preferisce allora seguire il proprio spirito avventuroso, esplorando territori e isole vicine, insieme all'immancabile Keller. Assapora un'ebbra libertà che lo avvicina sempre più alla poetica di Rimbaud e dei simbolisti e che gli fa esclamare: «le *Illuminations* di Arturo Rimbaud sono forse più note in Italia che in Francia, ma certo sono più capite là che qui. Pensavo intensamente a Rimbaud siccome a noi ormai Gabriele D'Annunzio dista indietro da Rimbaud quanto Numa Pompilio da Augusto» [5: 59]. Non trascorre molto tempo dal termine dell'impresa fiumana che il giovane Comisso, poco adatto al percorso di studi accademici, si imbarca sul veliero "Il gioiello" del capitano Gamba, a bordo del quale solca l'Adriatico da Chioggia fino all'Istria e alla Dalmazia. Di queste seconde peregrinazioni marinaresche, come si accennava, resta una memoria più viva e dettagliata in *Gente di mare*, pubblicato nel 1928 per i tipi di Treves e che gli frutta il Premio Bagutta.
Per restare ancora su *Il porto dell'amore*, trattandosi di una raccolta che ha conosciuto nel tempo diverse ripubblicazioni, l'obiettivo è quello di valorizzarla tenendo conto della dinamicità del testo: occorre, pertanto, elaborare un'edizione «in formato digitale interrogabile» e con «marcatura interoperabile», che permetta di verificare di volta in volta le aggiunte successive, favorendo così un «atto di critica delle varianti» [10: 49]. Un plausibile punto di riferimento potrebbe essere la *DSE* dei *Promessi Sposi* realizzata mediante la piattaforma *PhiloEditor* sviluppata dall'Università di Bologna, dove le varianti possono essere suddivise in base alle diverse metodologie correttorie (sostituzione di termini, inversione di parole, cancellazioni, interventi di punteggiatura, ecc.)[2]. Tuttavia, se il modello di *PhiloEditor* funziona benissimo nel caso

[1] Isabella Panfido scrive che «l'unica storia della narrativa di Comisso ruota intorno alla biografia e alla temperie emotiva che dalla esperienza deriva» [16: 145].
[2] https://projects.dharc.unibo.it/philoeditor/

manzoniano perché consente un confronto circoscritto a due edizioni, la 'ventisettana' e la 'quarantana', per cui l'edizione digitale ha il merito di generare un accesso privilegiato all'officina dello scrittore garantendo un confronto intuitivo e ben organizzato delle varianti – quindi adatto anche al lettore meno esperto [18] –, nel caso de *Il porto dell'amore* la situazione è diversa. Dal momento che le edizioni di cui tener conto sono in numero maggiore, si profilano due soluzioni: o si potrebbe optare per una selezione delle due edizioni più significative (ma risulterebbe un'operazione difficilmente giustificabile) o, ai fini di un'edizione critica digitale che miri all'esaustività, si dovrebbero prendere in considerazione tutte le edizioni, e semmai anche le prime stesure autografe. Per sviluppare questa seconda pista, sarebbe forse più opportuno ricorrere a una codifica del testo in XML/TEI da visualizzare con il software EVT, in maniera tale da fissare il testo base – quello corrispondente all'ultima volontà dell'autore, ovvero all'edizione Longanesi del 1959 – da proporre a un lettore non specialista, e poi tramite la modalità confronto consentire, su schermo bipartito, la visione in simultanea dei testi integrali delle altre edizioni, assunte come 'testimoni'. Tra la schiera dei testimoni dell'opera, oltre alla già citata edizione del 1924, andrebbero inserite quella del 1928, uscita per i tipi dei fratelli Ribet (edizione in cui il titolo del libro viene modificato in *Al vento dell'Adriatico*), e infine quella del 1953, ancora con lo stesso titolo, ma che rispetto alla precedente presenta una grande novità, ovvero l'accorpamento del *Porto dell'amore* e di *Gente di mare* (effettuato in virtù della contiguità tematica e stilistica delle due raccolte). Vale la pena ricordare che nel meridiano Mondadori dedicato a Comisso, i curatori Nico Naldini e Rolando Damiani si soffermano soltanto su alcune delle varianti che accompagnano la ripubblicazione dell'opera, senza effettuare una comparazione integrale dei testi [4: 1640-6].

3. IL CENTRO DI DOCUMENTAZIONE DIGITALE GIOVANNI COMISSO

Grazie al lavoro dell'Associazione Amici di Comisso, pochi anni fa ha visto la luce il *Centro di Documentazione Digitale Giovanni Comisso*, un archivio digitale che attualmente consta di circa quattromila file e che mira a conservare tutto quanto pubblicato in vita *su* e *da* Comisso[3], cui si aggiungono alcune corrispondenze inedite, di cui si sta recentemente occupando Giacomo Carlesso per inquadrare meglio il periodo parigino [2]. Oltre che nella gran mole dei materiali conservati e nell'intuitivo utilizzo delle opzioni di ricerca su cui si basa la piattaforma – la ricerca dei testi può avvenire inserendo il titolo del racconto, o in alternativa il nome del periodico su cui si ritiene pubblicato il racconto in questione o, persino, si può digitare l'incipit e ottenere il riscontro di tutte le sedi in cui compare la porzione testuale digitata – il vero punto di forza di questo archivio risiede, soprattutto, nella possibilità di rinvenire i testi in tutte le edizioni note, cioè di seguirne parallelamente la storia editoriale e redazionale. In altre parole, in questo *repository* si possono leggere i testi direttamente sulle sedi dove sono stati pubblicati dall'autore: tanto su giornali e riviste (sono state digitalizzate esclusivamente le colonne su cui estendono i testi comissiani), quanto poi nella forma licenziata per eventuali edizioni in volume (che sono state digitalizzate per intero). In questo modo, pur essendo per lo più esclusi i materiali inediti e quindi risulti impossibile venire a conoscenza di eventuali varianti rispetto ai manoscritti o dattiloscritti autografi, l'archivio digitale risulta uno strumento prezioso per il filologo che voglia confrontare le diverse stesure edite di un racconto, in cerca delle varianti d'autore che accompagnano il passaggio da una sede editoriale all'altra (vd. Fig.1).

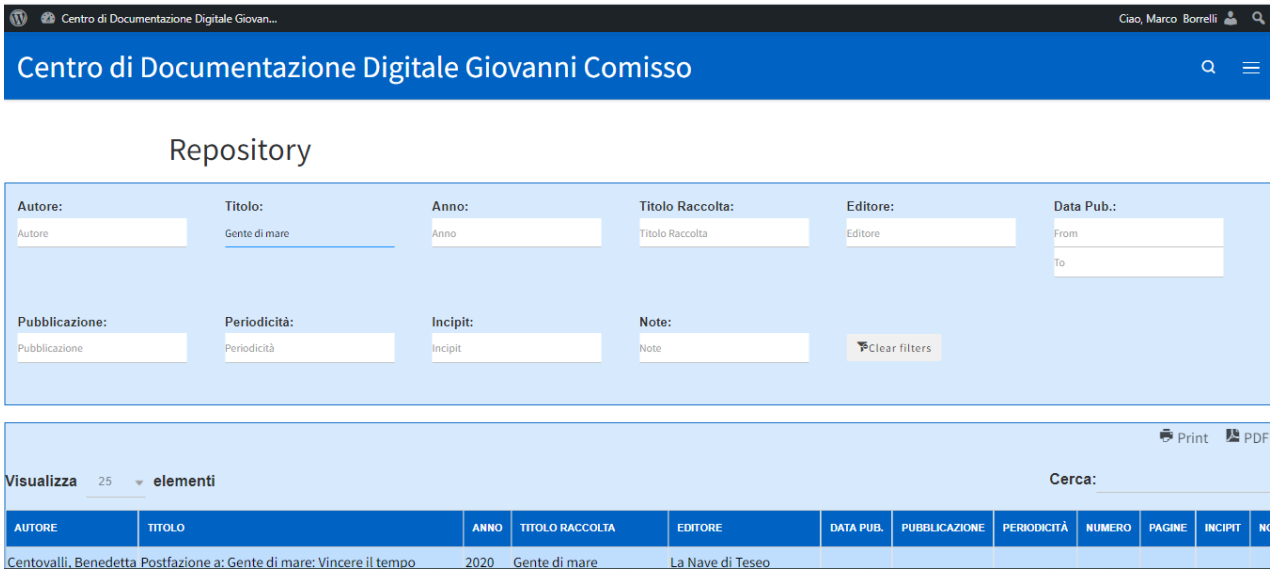


Figura 1. Gli strumenti di ricerca del Centro di Documentazione Digitale Giovanni Comisso.

[3] https://www.premiocomisso.it/archivio/repository-cdd/. Questo *repository* è una risorsa utile anche per i testi giornalistici [19].

Nonostante le grandi risorse messe a disposizione, il *Centro di Documentazione Digitale* presenta al momento una criticità: non essendo una piattaforma ad accesso libero – ma resta uno strumento a disposizione degli studiosi di Comisso, previa richiesta, accettazione e iscrizione da parte dell'Associazione – i testi digitalizzati non si prestano a un utilizzo interoperabile (vd. Fig. 2). È un aspetto su cui occorre lavorare, perché in vista di una futura edizione critica digitale, sarebbe un valore aggiunto ottenere delle licenze che consentano di inserire nella DSE dei *LOD*, così da creare una rete di dati e informazioni più facilmente interrogabile, all'insegna dei principi della FAIRness (Findability, Accessibility, Interoperability, Reuse) [20]. Si potrebbe fare ricorso a un link diretto, inserendo nelle note al testo l'URL relativo a un'immagine o a un testo, ma sarebbe auspicabile, per rendere l'operazione più solida e sistematica, ricorrere a immagini con framework IIIF, così da poterle visualizzare simultaneamente sullo schermo accanto al testo e renderle disponibili per ulteriori progetti e ricerche. Tuttavia, per quanto concerne questo aspetto, a onore del vero si deve dire che lo sforzo del singolo filologo non è sufficiente. Per bypassare i problemi legati al copyright e a operazioni digitali di vasta portata, che chiamano in causa competenze trasversali, è necessario lavorare in team; solo all'interno di progetti condivisi è possibile trovare di volta in volta delle soluzioni originali rispetto alla complessità delle sfide [12: 70-81].

L'EGO DEL PIAVE — Pag. III.

Chiacchiere con una Signora

La Signora m'invitò a passare nella cucina, dove una serva stava tutta intenta a disporre su d'una panca i piatti appena lavati. Ci si sedette a breve distanza e rimanemmo entrambi come preoccupati dei discorsi da fare per evitare la noia dell'attesa. La serva intanto già aveva alzato il volto per guardarmi. Un volto sfinito e rientrante in se stesso, come un fiore che ormai abbia concesso alle api tutto il suo piacere possibile. Pure un abito accuratissimo copriva il suo corpo, che ripugnava a immaginarlo. Una cravattina di velluto nero, spiccava con malizia sul verdone della blusa e le maniche rimboccate per bene, assieme a quelle bianchissime della camicia, lasciavano vedere le braccia aride come di terra cotta. I capelli tinti e ravviati diritti su dalla fronte, formavano una corona rigonfia che supposi contenesse tutti i suoi capelli morti, raccolti dal pettine.

Ancora non so spiegarmi come mai col desiderio di piacevoli forme che avevo potessi, non solo tollerare una così trista presenza, ma insistervi sopra con l'attenzione dei miei occhi.

La Signora mi distolse con un'intonazione animosa a parlarmi di spettacoli lirici teatrali. Non l'aveva determinata a questo tema il mio aspetto d'appassionato melodico, che quel giorno avevo, ma la una mattina d'inverno l'avesse trovato tutto raffrappito accanto all'abbaino, sotto una tegola, come l'avesse nutrito con pazienza proprio come un bambino ammalato e come fosse riescita ad ammaestrarlo. E qui mi parlò delle sue prodezze.

Mi indicò una cagnetta bianchiccia, accucciata su d'una sedia, in ombra, e della quale non mi ero accorto.

— Bisogna vedere, disse, le scene di gelosia che fanno tra loro queste due bestiole. Stia a vedere.

E incominciò: — Roma, vieni qui; Roma vieni dalla mamma, vieni poverina dalla mamma che ti dà la poppa, che ti fa bene alla pancetta.

Roma, alzò il muso di tra le zampe sottili, e accortasi che la padrona fingeva, non si mosse e rimase a guardarla.

Allora, la serva, fattasi innanzi, si curvò ad accarezzarla, il colombo subito che se ne accorse, alzò il capino verso la donna, mettendosi a tubare con arroganza, e a fare la ruota con le ali alzate.

— Vede? Vede come s'arrabbia? — mi disse la Signora. Aspetti, aspetti! Ora vedrà che anderà a beccarla ai piedi...

Difatti, mentre la serva continuava le sue moine verso la cagnetta, il colombo si appressò di scatto ad uno dei suoi piedi, fermo sul pavimento, lo colpì col becco e subito si rifrasse.

Figura 2. Racconto Chiacchiere con una signora su "Eco del Piave": digitalizzazione realizzata dal Centro di Documentazione

Al momento, data la complessità della questione, che si tratti di allestire una DSE de *Il porto dell'amore* o di *Gente di mare*, sarebbe già un significativo passo in avanti, accanto alla presentazione della storia variantistica, la realizzazione di un'ulteriore nota al testo – magari ricorrendo al tag `<note>` (nel caso di EVT la nota viene segnalata da un pallino accanto alla porzione testuale interessata) – che indichi le relazioni vigenti tra i racconti e le prime esperienze biografiche e giornalistiche dell'autore trevigiano. Infatti, gli articoli scritti durante l'esperienza di Fiume e pubblicati per le colonne de "Il Risorgimento", "Camicia nera" e "Yoga" sono fondamentali per comprendere la temperie culturale in cui si muove Comisso, nonché i suoi interessi eclettici che vanno dall'arte e l'architettura all'estetica del paesaggio naturale; interessi che sono legati a doppio filo con la scrittura narrativa.

4. PER UN'EDIZIONE CRITICA DIGITALE DI *GENTE DI MARE*

Per l'edizione critica digitale di *Gente di mare*, su cui ho portato leggermente più avanti il lavoro, si è preferito intraprendere un'operazione di codifica finalizzata a una visualizzazione in EVT, perché d'accordo con quanto sostiene Elena Pierazzo, si ritiene che il software prodotto da Roberto Rosselli Del Turco [14] risponda adeguatamente alle esigenze di una DSE che, con riferimento al mondo della moda, voglia essere «accessibile e portabile», come «l'abbigliamento del *prêt-à-porter*» [17: 8]. Da un punto di vista filologico, per la ricostruzione della storia variantistica dei testi della raccolta, che analogamente a quelli de *Il porto dell'amore* sono circolati alla spicciolata su diversi giornali e periodici prima di essere riuniti in volume, ho potuto usufruire, oltre che delle digitalizzazioni recuperate dal *Centro di Documentazione*, di foto scattate personalmente presso il Fondo Comisso, dove è stato possibile addirittura, per alcuni racconti, recuperare le prime stesure manoscritte inedite. Dal punto di vista della resa digitale, di conseguenza, si sta allestendo un'edizione che permetta all'utente di visualizzare il testo relativo all'ultima volontà dell'autore, con la presenza di una nota al testo e di un apparato in cui sono trascritte le lezioni degli altri testimoni (vd. Fig. 3).

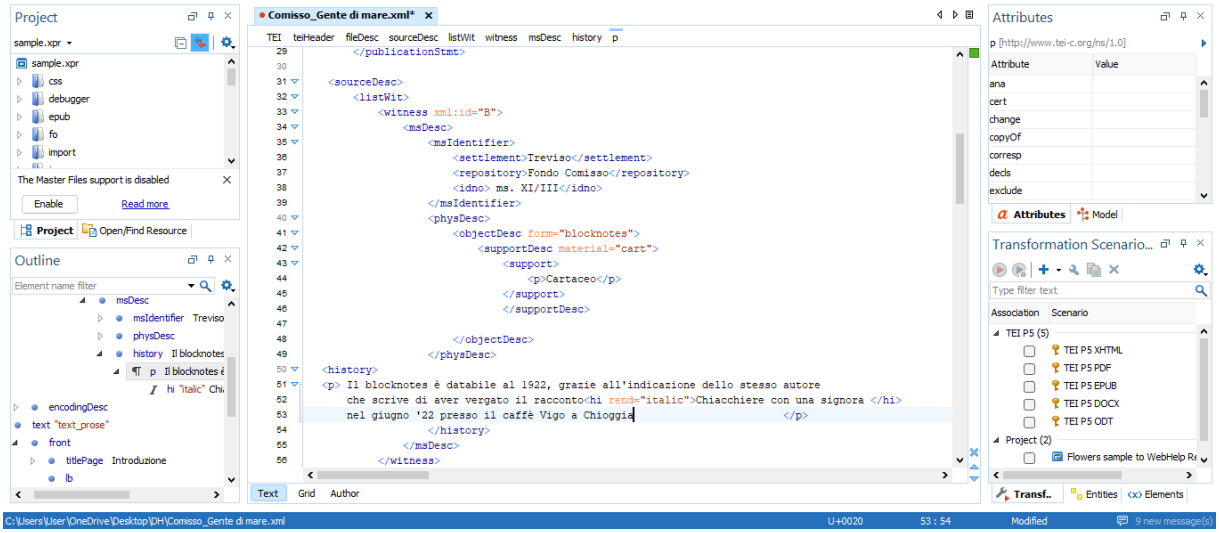


Figura 3. Esempio di descrizione del testimone "B": block notes consultato presso il Fondo Comisso di Treviso.

Trattandosi di un caso di filologia genetica, nell'apparato si registrano le varianti presenti nelle edizioni precedenti a quella corrispondente all'ultima volontà autoriale, e si aggiungono altresì, laddove rinvenuti, i testi dei racconti nella loro stesura originaria, che differiscono in maniera sostanziale da quelli successivamente pubblicati. Con questa edizione digitale che si propone di essere tale anche «nell'ambito della filologia classica» [13: 26] si vorrebbe, in ultima istanza, coniugare l'esigenza di restituire un testo filologicamente attendibile e, al tempo stesso, dare al lettore l'impressione che si è dinanzi a testi in movimento, anche perché, nel corso degli anni, l'autore cambia perfino la struttura dell'opera, rimpinguandola con nuovi racconti nel passaggio da un'edizione all'altra.

Questa DSE andrebbe ad inserirsi, inoltre, in un preciso discorso ermeneutico, inerente allo stile di Comisso che troppo spesso è stato tacciato di superficialità, quasi che l'autore, preso dall'unico scopo di registrare la vita in presa diretta, non avvertisse il bisogno di lavorare di lima sulla grammatica e sulla sintassi. Seppur complicato, sarebbe interessante integrare alcune immagini scattate in archivio, anche solo per mostrare la varietà dei supporti cartacei di cui l'autore si serve per annotare ciò che i suoi sensi percepiscono dal paesaggio circostante: perché Comisso si rivela scrittore d'istinto non nella volontà di lasciare il testo così come scaturisce per la prima volta dalla sua penna, bensì nell'urgenza di annotare tutto ciò che proviene dalla natura e dall'uomo. Per appuntare impressioni e riflessioni, o addirittura per scrivere bozze di racconti, Comisso si serve di taccuini di vario formato, ma non mancano quaderni, block notes, fogli sparsi o buste per lettere: si ha la sensazione che qualunque tipo di supporto venga considerato indispensabile per contrastare la fugacità del tempo (vd. Fig. 4).

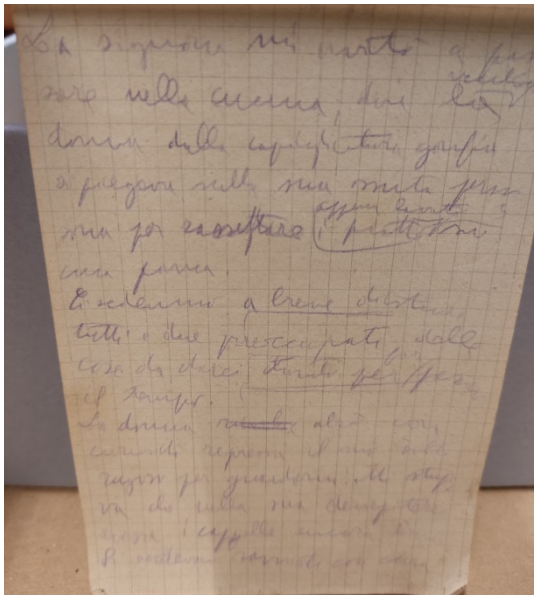


Figura 4. Il 'manoscritto' autografo che riporta la prima stesura di Chiacchiere con una signora

Si è già detto che la raccolta *Gente di mare* cresce col tempo, in quanto al nucleo originario si aggiungono altri testi non ideati all'origine per far parte del volume (vengono 'sforbiciati' da altre accolte) [3: 275-285]. Questa mescolanza fa emergere una duplice questione: tematica e stilistica. Le prime prose che compongono l'opera, che fotografano la vita di bordo e le ore trascorse insieme al capitano e all'equipaggio de "Il gioiello", piuttosto che alla narrativa breve sembrano afferire alla tradizione francese del *poème en prose*. Fernando Bandini è il primo a mettere in evidenza la natura poetica di questi racconti di mare, scorgendo delle importanti analogie tra le pagine di Rimbaud e quelle di Comisso [1: 63-67]; per di più, scavando nella cosiddetta *Preistoria di Comisso*, il critico fa notare che i primi passi letterari dell'autore muovevano proprio verso la poesia, con la produzione di componimenti vicini allo stile dei frammentisti della "Voce". Oltre che per i pregi letterari, questa raccolta risulta estremamente moderna perché si inserisce bene nelle coordinate storiografiche delineate da Egidio Ivetic, dove lo studioso integra la civiltà dell'Adriatico nel più ampio contesto della storia del *Mare nostrum*, rivendicando così le profonde radici mediterranee della regione veneta, la cui identità attuale sembra, invece, sempre più proiettata verso una dimensione industriale [11]. *Gente di mare* attira subito l'attenzione di Carlo Emilio Gadda, che definisce Comisso uno scrittore «potentemente dotato di tutte le impressioni sensoriali» [4: 1636]. Quest'opera è sottesa da una duplice e, per così dire, ossimorica volontà: quella dell'ozio, dove prevale il desiderio dell'autore di lasciarsi andare a rotte insolite per evadere dalla realtà contemporanea, e quella di matrice etnoantropologica – di cui forse serberà memoria Pasolini – che si dispiega nella ricerca di luoghi vergini, nei quali poter ritrovare costumi e tradizioni simili a quelli dei contadini e dei pescatori veneti. Così, a distanza di ormai quasi un secolo, la ricerca di Comisso si pone ancora come un legittimo punto di partenza per abbracciare un'idea di Mediterraneo nella quale le dinamiche dell'incontro culturale prevalgano su quelle dello scontro e dell'incomprensione reciproca.

Per l'edizione critica digitale dell'opera, si dispone dunque facilmente delle prime pubblicazioni dei racconti, tutti consultabili tra i materiali digitalizzati dal *Centro di Documentazione Digitale*. Per lo più compaiono, per la prima volta, essenzialmente su quattro giornali: nello specifico, su "L'eco del Piave", "Il Convegno", "Il Quindicinale" e "Solaria". Il confronto tra le varie edizioni dell'opera va allora esteso sicuramente a questi racconti apparsi alla 'spicciolata', ma pure, laddove sia possibile rinvenirle negli archivi, alle prime stesure degli stessi (vd. Fig. 5). Tra le carte dello scrittore custodite presso la Biblioteca comunale di Treviso è stato possibile ritrovare, per esempio, la prima stesura a penna (del 1922), su un block notes e con inchiostro appena leggibile, del racconto *Chiacchiere con una signora*, edito su "L'eco del Piave" nell'agosto 1925 e poi ancora, l'anno successivo, sul primo numero di gennaio de "Il quindicinale".

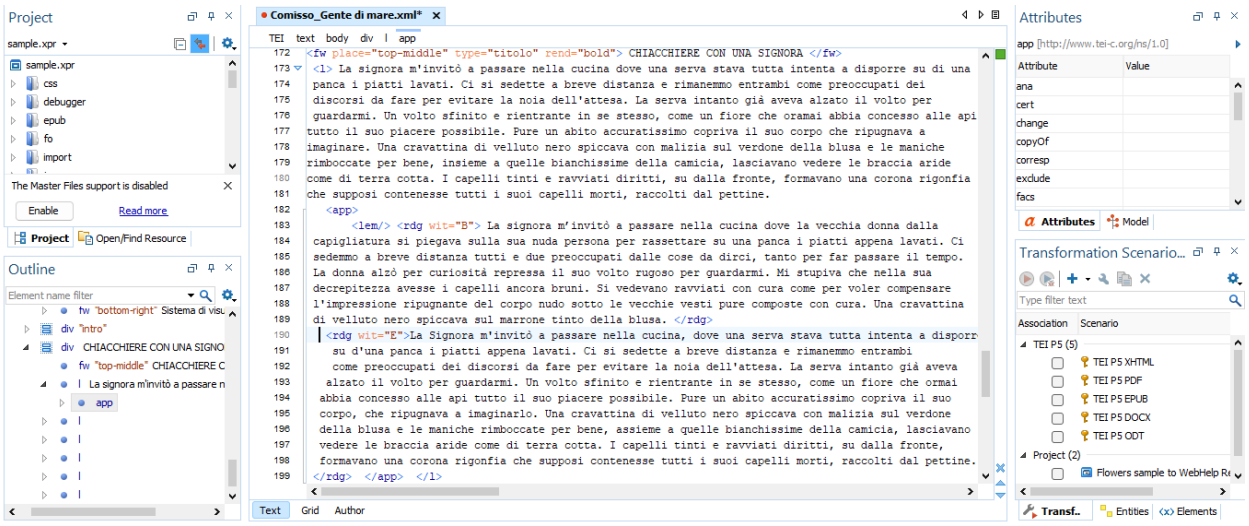


Figura 5. Le varianti riscontrabili nel racconto pubblicato su "Eco del Piave" e nella stesura autografa su block notes

In conclusione, il riferimento esemplificativo a questo racconto, che potrebbe non essere l'unico (ho individuato già dei dattiloscritti con correzioni autografe, mentre per altre stesure manoscritte andrebbero indagati gli archivi dei direttori dei giornali, perché quello custodito alla biblioteca di Treviso risulta stranamente lacunoso), chiarisce bene la necessità di un'edizione critica digitale che consenta una più facile mappatura filologica dell'opera di Comisso. Non sarebbe lavoro vano inseguire la storia editoriale e variantistica di ciascun racconto, a partire dalla grafia illeggibile dei block notes fino all'ultima edizione di *Gente di mare* uscita per Longanesi nel 1966.

BIBLIOGRAFIA

[1] Bandini, Fernando. «Preistoria di Comisso». In *Giovanni Comisso*, a cura di Giorgio Pullini, 59–71. Firenze: Olschki, 1983.

[2] Carlesso, Giacomo. «Una continua invenzione. La Parigi di Comisso». In *Giovanni Comisso. Uno scrittore trevigiano e il suo archivio. Atti della giornata di studio. Treviso, 28 maggio 2022.*, 111–36. Crocetta del Montello: Antiga, 2023.

[3] Comisso, Giovanni. *Gente di mare*. Milano: La nave di Teseo, 2020.

[4] Comisso, Giovanni. *Opere*. A cura di Rolando Damiani e Nico Naldini. Milano: Mondadori, 2002.

[5] Comisso, Giovanni. *Solstizio metafisico*. A cura di Annalisa Colusso. Padova: Il poligrafo, 1999.

[6] Comisso, Giovanni. *Viaggi nell'Italia perduta*. A cura di Nicola De Cilia. Roma: Edizione dell'asino, 2017.

[7] De Cilia, Nicola. *Geografie di Comisso. Cronaca di un viaggio letterario*. A cura di Gregorio Maria. Vicenza: Ronzani, 2019.

[8] De Cilia, Nicola. *Giovanni Comisso. Un invito alla lettura*. Udine: Digressioni, 2021.

[9] Demattè, Francesca. *Giovanni Comisso e Mario Botter nella Fiume di D'Annunzio 1919-1921*. Crocetta del Montello: Grafiche Antiga, 2021.

[10] Italia, Paola. «Per una critica delle varianti digitale». In *Moving texts. Filologie e digitale*, 41–58. Napoli: UniorPress, 2023.

[11] Ivetic, Egidio. *Il grande racconto del Mediterraneo*. Bologna: Il Mulino, 2022.

[12] Mancinelli, Tiziana, e Elena Pierazzo. *Che cosa è un'edizione scientifica digitale*. Roma: Carocci, 2020.

[13] Michelone, Francesca. «L'edizione critica tra digitale e stampa: riflessioni metodologiche». *Umanistica Digitale* 10 (2021): 25–48.

[14] Monella, Paolo, e Roberto Rosselli Del Turco. «Extending the DSE: LOD Support and TEI/IIIF Integration in EVT». In *La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica. Atti del IX Convegno Annuale dell'AIUCD*, 148–55. Milano: Università Cattolica del Sacro Cuore, 2020.

[15] Naldini, Nico. *Vita di Giovanni Comisso*. Napoli: L'Ancora del Mediterraneo, 2002.

[16] Panfido, Isabella. «Il prosimetro in Comisso, un'ipotesi di rilettura». In *Giovanni Comisso nel tempo. Atti della giornata di studi. Treviso, 7 novembre 2019.*, 145–49. Treviso: Associazione Amici di Giovanni Comisso, 2020.

[17] Pierazzo, Elena. «Quale infrastruttura per le edizioni digitali? Dalla tecnologia all'etica». *Textual Cultures* XII, fasc. 2 (2019): 5–17.

[18] Russo, Ersilia. «Manzoni digitale. Philoeditor tra filologia e didattica». *Griseldaonline* 20, fasc. 2 (2021): 162–72.

[19] Sandrini, Giuseppe. *Per Giovanni Comisso. Critico, editore, giornalista*. Sommacampagna: Cierre, 2023.

[20] Tomasi, Francesca. «Web semantico, Linked Data e archivi. Metodologie e strumenti per la rappresentazione della conoscenza». *Comunicazione storica. Tecnologie, linguaggi e culture*, 2021, 237–55.

[21] Tonini, Paolo. *Il porto dell'amore. Rivolta e poesia di Fiume dannunziana*. Gussago: Arengario, 2020.

[22] Urettini, Luigi. *Giovanni Comisso: un provinciale in fuga*. Treviso: Cierre, 2009.

ARCHIVI E MUSEI DIGITALI PER IL PATRIMONIO CULTURALE

A Workflow for GLAM Metadata Crosswalk

Arianna Moretti[1], Ivan Heibi[2], Silvio Peroni[3].
[1]University of Bologna, Italy – arianna.moretti4@unibo.it
[2]University of Bologna, Italy – ivan.heibi2@unibo.it
[3]University of Bologna, Italy – silvio.peroni@unibo.it

ABSTRACT

The acquisition of physical artifacts not only involves transferring existing information into the digital ecosystem but also generates information as a process itself, underscoring the importance of meticulous management of FAIR data and metadata. In addition, the diversity of objects within the cultural heritage domain is reflected in a multitude of descriptive models. The digitization process expands the opportunities for exchange and joint utilization, granted that the descriptive schemas are made interoperable in advance. To achieve this goal, we propose a replicable workflow for metadata schema crosswalks that facilitates the preservation and accessibility of cultural heritage in the digital ecosystem. This work presents a methodology for metadata generation and management in the case study of the digital twin of the temporary exhibition "The Other Renaissance – Ulisse Aldrovandi and the Wonders of the World". The workflow delineates a systematic, step-by-step transformation of tabular data into RDF format, to enhance Linked Open Data. The methodology adopts the RDF Mapping Language (RML) technology for converting data to RDF with a human contribution involvement. This last aspect entails an interaction between digital humanists and domain experts through surveys leading to the abstraction and reformulation of domain-specific knowledge, to be exploited in the process of formalizing and converting information.

KEYWORDS

Schema Crosswalk; Workflow; RDF; Digital Twin; Cultural Heritage.

1. INTRODUCTION

The digitization process offers the advantage of extending access to cultural heritage geographically and chronologically [21, 22]. Additionally, to foster preservation in the digital ecosystem, the acquisition of artifacts must be accompanied by FAIR metadata creation, management, and maintenance [17, 24]. Temporary exhibitions introduce further elements of interest and complexity in formalizing GLAM (Galleries, Libraries, Archives, and Museums) data [15], as they often feature heterogeneous objects from multiple sources, potentially described according to different representation models. Thus, conversions between formats and *schema crosswalks* between data models are needed to exploit data jointly for specific purposes.

The term *schema crosswalk* refers to the mapping between conceptualization systems that describe at least partially overlapping domains, intending to identify points of contact and divergence to facilitate data exchange. In this context, mappings between schemas for metadata management in the description of digital objects have been created [20], sometimes accompanied by invitations to best practices, such as harmonizing existing schemas and using recommended data types to avoid semantic loss [7].

In this paper, we introduce a workflow for simplifying the creation and, thus, reproducibility of schema crosswalks exploiting digital humanists' and domain experts' contributions. This approach seeks to balance the reuse of general components with solutions developed ad hoc for the case study, i.e. the creation of the digital twin of the temporary exhibition "The Other Renaissance - Ulisse Aldrovandi and the Wonders of the World"[1] [3, 4]. Specifically, we focus on providing a process for converting data in tabular form, describing the objects included in the exhibition and the digitization process for creating their digital replicas, into the Resource Description Framework (RDF) format. The goal is to enable the use of Semantic Web technologies on the content of exhibition material descriptions and acquisition metadata, formulated as Linked Open Data (LOD).

The paper includes a Literature Review section, providing an overview of several approaches for exploiting structured data as RDF. The Case Study of the temporary exhibition "The Other Renaissance – Ulisse Aldrovandi and the Wonders of the World" is then introduced, to illustrate the context of application of the workflow. The Methodology section presents the procedure step-by-step, providing reasons for the adopted choices. In the Future Developments, we introduce some potential and desirable evolutions of the approach, to deliver a more complete and reusable research product. In the Conclusions, we summarise the content of the paper to highlight the key aspects of the proposed workflow.

[1] https://site.unibo.it/aldrovandi500/en/mostra-l-altro-rinascimento

2. LITERATURE REVIEW

The conditions to perform schema crosswalks are closely tied to a system's interoperability, meant as the property of a data model to exchange information seamlessly. On this topic, the EOSC Executive Board FAIR Working Group's Interoperability Task Force produced a report [7] outlining directives for facilitating data exchange, with a focus on the FAIR principle of interoperability across technological, semantic, organizational, legal, and syntactic levels. The document provides a crosswalk of data models, controlled vocabularies, and aggregators' guidelines, aiming to harmonize existing patterns for mapping and prevent semantic reductions due to metadata quality loss. Similarly, Milan Ojsteršek published a noteworthy crosswalk on Zenodo that maps the basic properties of widely used vocabularies [20], creating a comprehensive knowledge base for future developments in the metadata exchange domain.

A practical situation in which it is required to perform metadata crosswalks is the digital objects integration within collectors. Such activity often involves adapting data and metadata to meet the needed target formats and models, and users typically have to face challenges based on platform documentation clarity, familiarity with target formats, model knowledge, and technical tool proficiency. Indeed, many services share the common drawback of limited declarative support to the user for performing the crosswalk. Coherently, among the causes for metadata quality loss in the description of digital objects, the EOSC Interoperability Framework Report mentions: (1) mismatches between the updates of artifacts, vocabularies, metadata management software, and administered data; (2) compelled metadata conversion adjustments, and (3) unsupervised inclusion of freely structured metadata in the content aggregators [7]. Thus, a workflow for guiding the crosswalk process for uploading and updating data, and facilitating the selection of conceptual and technical tools for information exchange between formats, would be beneficial.

More in detail, in the Semantic Web domain, it is necessary to transform other formats into RDF or to query them as if they were RDF-structured. Some direct conversion tools that allow various formats to be accessed as if they were formalized in RDF enable such conversion but do not aim to create a unique abstraction for format management. Examples of this category are Any23[2], JSON2RDF[3], and CSV2RDF[4] [9].

Other approaches also offer an abstraction over the particular format used to define data. Such tools include D2RQ [6], a system for accessing the contents of relational databases as RDF graphs, thanks to a server in which a conversion of SPARQL query to SQL is implemented. On the other hand, Triplify [2] was among the first software components to be integrated into other applications for converting data from relational databases to RDF. In this technology, the semantic content is maintained by mapping HTTP-URI requests to SQL queries.

Other solutions include the RDF Mapping Language (RML) [11], a generic mapping language with customizable rules independent of specific implementations. RML was a forerunner in the reengineering approaches of formats according to the RDF model as a consequence of the use of SPARQL, also enabling the querying of relational databases. This technology was selected for the workflow step concerning the conversion to RDF for its accurate documentation and the wide set of open-source tools provided by the developers to facilitate the pipeline execution. In addition to that, the Aldrovandi temporary exhibition case study foresaw the necessity to convert information stored in CSV tables with non-trivial structures into RDF serializations. This circumstance motivated the adoption of RML also because of the possibility of extending the mapping rules through the definition of custom functions in Java and Turtle to potentially meet any specific need concerning the conversion process. A valuable alternative taken into account was Facade-x [8]: a generic meta-model that allows querying resources as if they were structured in RDF using wrappers, without extending the SPARQL syntax. It is concretely implemented in SPARQL Anything, a reengineering system that facilitates querying any structured data using SPARQL and creating knowledge graphs without requiring specific skills in a particular mapping language or an in-depth understanding of formats. However, the process implies automated reengineering to the target meta-model and domain knowledge reframing by an RDF and SPARQL expert. At a technical level, it implements a set of transformers mapped to various media types, extendable with Java classes. In addition to the above-mentioned tools, SPARQL Generate [18] should be mentioned as a declarative transformation language extending the SPARQL syntax for generating RDF graphs or textual streams, that can be further integrated with mediating query languages such as XPath[5] to handle new sources.

[2] https://any23.apache.org/
[3] https://github.com/AtomGraph/JSON2RDF
[4] https://clarkparsia.github.io/csv2rdf/
[5] https://www.w3.org/TR/xpath/

3. CASE STUDY

The temporary exhibition "The Other Renaissance - Ulisse Aldrovandi and the Wonders of the World", hosted at the Palazzo Poggi Museum in Bologna for six months starting from December 2022, showcased hundreds of artifacts belonging to the naturalist Ulisse Aldrovandi (1522-1605). To acknowledge its cultural significance, an experimental approach to save temporary exhibitions was undertaken [3].

In this context, a methodology for collecting, formalizing, managing, and exporting FAIR cultural heritage metadata was defined. More in detail, the specificity of this case study was closely related to the nature of the project of digitizing a temporary exhibition, envisaging the necessity to manage both the information derived from museum descriptions concerning the exhibited objects and the data generated during the digital acquisition process.

The metadata generation and management process was structured in steps. The first stage involved the tabular formalization of information on the museum objects and the acquisition processes, for the creation of two input datasets. Specifically, the first of these was structured based on metadata inferred from museum captions and - where possible - from catalogs' data. The result was a table with the following fields: Identification Number, Linked Identification Number, Relationship, Exhibition Room, Caption, Consistency, Documentary Typology, Technique, Reproduction Typology in Exhibition, Subjects, Original Title, Museum Title, English Title, Date, Discoverer, Author, Translator, Illustrator, Engraver, Publisher, Place of Publication, Museum Preparer, Commissioner, Parent Work Typology, Parent Work Title, Volume, Collection, Conservator Entity, Location of Conservation, Placement, Source, Digital Image, and Iconography. The table concerning the acquisition process, on the other hand, had a more complex structure, involving fields with various internal subdivisions. At the most general level, it provided data about Identification Numbers, Objects, Showcases, Captions, Current Status, Link, Notes, Acquisition, Processing, Modeling, Optimisation, Export, Metadata, and Upload.

In the meantime, to semantically describe the collected data, an Application Profile named CHAD-AP[6] was conceptualized [5] based on CIDOC CRM and CRM Digital, with a module for each of the two tables. Once the tables had been exported to CSV, they could be converted into RDF. From this point on, the traceability of the source information is maintained [19] by exploiting the OpenCitations data model [10]. Finally, the process ends with a quantitative analysis, visualization, and narration of the data using the tool MELODY [23].

This work focuses on the aspects related to converting the source dataset into RDF serializations. Our proposal (i.e., workflow) integrates software tools and human contributions from digital humanists and domain experts (in this case, museum curators).

4. METHODOLOGY

We present a workflow (see Figure 1) for converting GLAM metadata collections available in tabular form to N-Triples RDF serializations, highlighting the human role in its design and execution. At a technical level, the procedure relies on LimeSurvey[7], parsers for JSON and YAML[8], and PYRML[9] for producing the output collection from the input dataset, exploiting RML mapping rules.

[6] /DH.arc. "Dharc-Org/Chad-Ap." Jupyter Notebook. 2024. Reprint, /DH.arc, April 17, 2024. https://github.com/dharc-org/chad-ap.

[7] LimeSurvey. "LimeSurvey," January 26, 2024. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:c36ac88e2358050ad4d00ca457aef1ce6d0a3180;origin=https://github.com/LimeSurvey/LimeSurvey;visit=swh:1:snp:cf6c3f1c990cfc9f0f91e295eace369de35df0a3;anchor=swh:1:rev:a8b1f5d8b5b45d3e5a9eba4fbeb92ecd2398cdb2.

[8] RMLio. "YARRRML Parser," November 23, 2023. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:b2a545fb1ca9a8e42898db5413bdb2cd98af909b;origin=https://github.com/RMLio/yarrrmlparser;visit=swh:1:snp:9d7369e64633ab91c03d48938ea542cb38a81366;anchor=swh:1:rev:52e74c0bf4a554ac0d5e376378c50565803a67bb.

[9] Nuzzolese, Andrea Giovanni, and Giulio Settanta. "pyRML," 2024. https://archive.softwareheritage.org/swh:1:dir:7c7daa4e9e44e40be9b69b90130ed2c675df0cf6;origin=https://github.com/anuzzolese/pyrml;visit=swh:1:snp:7434d2226b78603ce653fffadf575387e26419cd;anchor=swh:1:rev:cd42ca3a92297c3527b537e24f42202a0295a4a7.

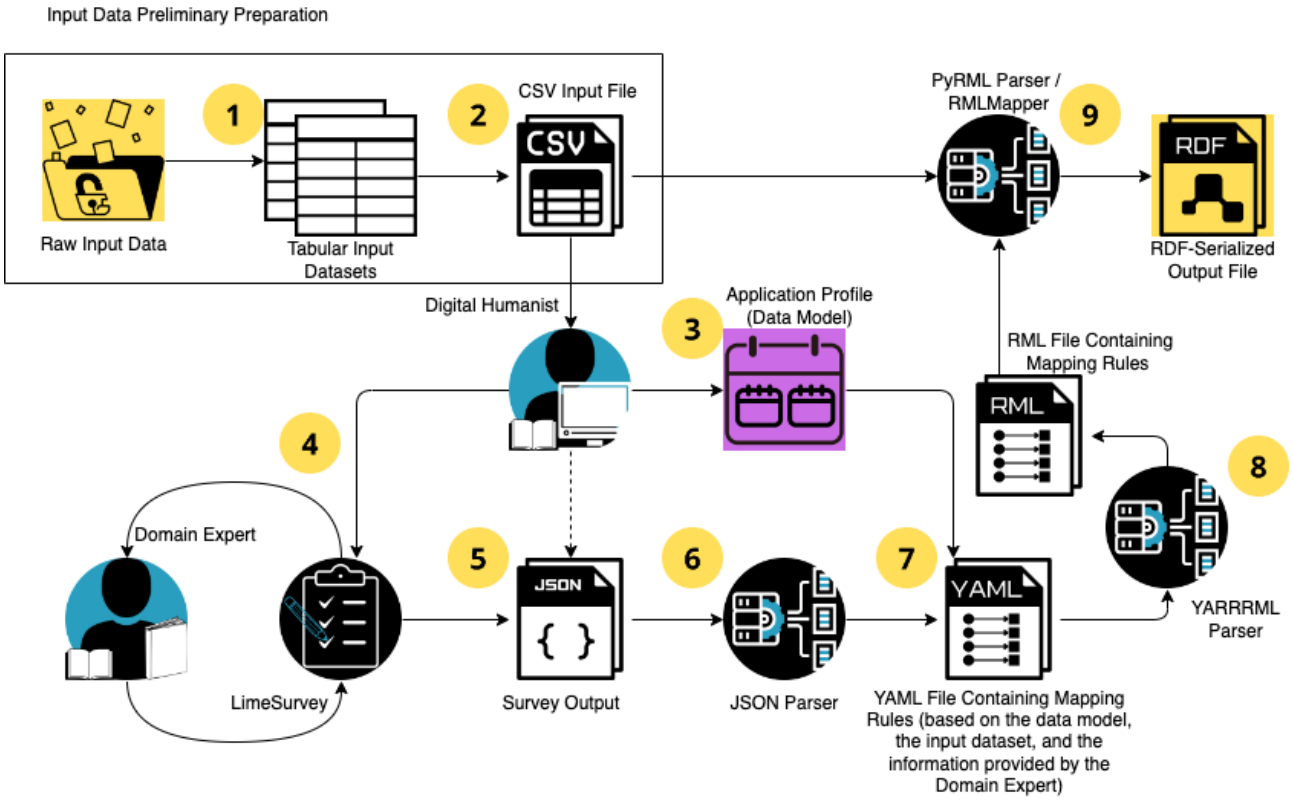


Figure 1. A workflow for CSV to RDF schema crosswalk of GLAM data involving human contributions

Since the selection of the conversion technology contributes to determining the possibilities of querying and exploiting the generated data, as well as the structure of the workflow and the involvement of domain experts, the choice for the mediation technology has fallen on RML. In addition to the aforementioned customizability, among the advantages of this technology is the opportunity to express the basic domain knowledge in configuration files, without extending software or dealing with the complexities of direct use of technical tools.

The process follows a series of well-structured steps (see Fig. 1). First of all, (1) the information obtained from the museum descriptions and the digital acquisition process is formalized in two separate tabular datasets, (2) that are then exported as CSV files. Unless an existing data model was reused to structure the input tables, during the first stage (3) the digital humanist examines the datasets and starts formulating an appropriate Application Profile, including one module for each of the datasets to describe, to exhaustively represent the gathered data. The obtained data model should have reached a stable version before starting the formulation of the RML rules. Afterward, (4) the digital humanist, with consolidated knowledge in the use of RML and associated technologies, formulates a LimeSurvey questionnaire to be presented to the domain expert, intending to guide formalizing their knowledge. This step aims to verify that the input information was correctly structured in the tables (i.e., that the museum descriptions were seamlessly interpreted) and that the data model does allow to convey all the informative content of the datasets. In addition to that, the questionnaire is an opportunity to pose additional questions to the domain experts, both to possibly improve the data model through unstructured text feedback and to gather additional preferences on the conversion output, where no formal constraints are posed by the data model or customization is needed. For example, as for the most recent model update, CHAD-AP imposes a constraint on the use of controlled vocabularies. Regarding the creation techniques, the model defines an Expression (*frbroo:F2_Expression*) as a reference to the intellectual content of the object, which is generated through a creation event (*frbroo:F28_Expression_Creation*), that uses various creation techniques (*crm:P32_used_general_technique*), whose domain is an Activity (*crm:E7_Activity*) and whose range is a Type (*crm:E55_Type*). The information to be transposed in RDF is in the "Technique" field of the CSV file of the object dataset and, in the case study, it was expressed as a textual string (e.g.: "drawing technique"). However, to make the RDF transposition fully compliant with the data model, we adopted the Art & Architecture Thesaurus[10] codes (e.g.: aat:300054196, "drawing (image-making)"), and thus ad hoc functions had to be defined for performing the conversion. Different case studies might involve specific modifications to the Application Profile to be efficiently reused and properly represent the input data. Therefore, a survey question on

[10] https://www.getty.edu/research/tools/vocabularies/aat/

adopting the same convention enables either use or don't the ad hoc conversion function. The following steps involve (5) exporting the questionnaire responses in JSON format to allow (6) the digital humanist to analyze the output and decide whether to review the data model and adjust the JSON file, in case the survey revealed further or unexpected information. (7) The rest of the data, along with the data model, is then used as input for the JSON parser to compile a YAML configuration file compliant with the syntax specification of the human-readable text-based representation for declarative mapping generation rules, named YARRRML [14]. At this stage of the process, (8) the obtained YAML file is used as input for the YARRRML parser, converting the configuration file into an RML Turtle serialized file. Eventually, (9) the input dataset and the RML configuration file are taken in input by the PyRML parser, to finally produce the RDF output dataset. As an alternative, this step can be performed by using the official Java RMLMapper[11].

It is worth noting that while it is possible to bypass the JSON-YAML conversion step by directly generating the RML file, it is recommended to follow the complete procedure since this helps limit the risk of errors by maintaining a gradual approach and providing an information formalization in a human-readable format.

The human contribution is condensed in the interaction between the domain expert and the digital humanist through the questionnaire formulation and compilation. In addition, questionnaires have been used to further optimize the process of formalizing domain information from experts who may lack coding skills. After evaluating options such as Microsoft Forms, SoGoSurvey, Google Forms, Typeform, SurveyMonkey, and Qualtrics, the choice fell to LimeSurvey. This tool, developed for academic and institutional purposes, is open source, freely accessible online, offers a widely usable free plan, can be configured to comply with GDPR, and allows the export of responses in various structured formats [16], a fundamental aspect for the subsequent conversion into RML configuration files. Further, with the perspective of adopting the same Application Profile to convert to RDF similarly structured collections, it ensures both the reuse of survey sections and the drafting of new custom-made questions, thus enabling the application of conversion patterns common for most metadata crosswalks and the expression of new requests to grasp the understanding of case-specific aspects. It is noteworthy that, while most of the information can be automatically formalized through a priori formulation of questions, new elements may emerge from the domain expert answers, causing the occasional intervention of the digital humanist on the data model. Furthermore, this human-centered step also implies verification of the JSON output's structural correctness before proceeding, in case of inconsistencies.

5. FUTURE DEVELOPMENTS

The presented case study allowed the definition of a reusable workflow that simplifies the creation of metadata schema crosswalks of GLAM data, from tabular to RDF format. However, considering the opportunities for joint use of variably structured cultural heritage data, a future development goal is to extend the workflow, making it potentially reusable for conversions between any-to-any formats. As an ambitious project, an intermediate milestone is proposed to achieve coverage in both input and output of widely used formats, including JSON, RDF, tabular, and XML, with their respective serializations.

Another potential evolution involves transitioning from a linear to an iterative workflow structure, where the human intervention limited to an initial phase is replaced by a "Human in the Loop" approach [1]. In this perspective, the domain expert is consulted again at the end of the process and participates in the evaluation of the output, guided in the formalization of their opinion through another questionnaire. In this way, the feedback received could be reintegrated as an additional input into the process, aiming to produce incremental versions of the output data until an optimal result is achieved.

Finally, the current workflow involves the direct and concatenated use of several software components in various programming languages. We argue that the reproducibility of the methodology could be enhanced by leveraging open-access tools for formalizing the procedure by providing examples of the execution phases and orchestrating the functioning of the technologies involved. In the first place, the steps of the procedure can be structured in an actionable workflow released on dedicated platforms that guarantee an adequate versioning system and free access to content, such as the Social Sciences & Humanities Open Marketplace[12] and Protocols[13]. The reproducibility potential could be further enhanced by tools that automatically execute the software components in a specific order, managing the inputs and outputs of intermediate steps. In this regard, MITAO [12, 13] represents a valid choice, ensuring declarative management automation

[11] RMLio. "RMLio/Rmlmapper-Java." Java. 2018. Reprint, RDF Mapping Language (RML), April 1, 2024. https://archive.softwareheritage.org/swh:1:dir:bc91cd466746b8e6ae65d183589501dfafeb5df2;origin=https://github.com/RMLio/rmlmapperjava;visit=swh:1:snp:c3deeafbc6826c89dac0b831992ba21f858e7874;anchor=swh:1:rev:f8d15d97efb9a30359b05f37a28328584fe62744.

[12] https://marketplace.sshopencloud.eu/

[13] https://www.protocols.io/

and process scalability. The software provides a user-friendly visual interface that facilitates users without programming skills to integrate data and tools in a customizable visual workflow and share the defined product. Additionally, the software can be extended and customized with the help of code developers (e.g., for integrating a specific tool for exporting LimeSurvey questionnaire results in JSON or converting it to YAML). As for the drawbacks, the current limitation of MITAO lies in its language-specific setup and it should, therefore, be extended to allow the combination of Python components with those in other programming languages, such as Java.

6. CONCLUSIONS

We presented a workflow for facilitating cultural heritage metadata conversions between different formats and models, fostering collaborative use, and leveraging Semantic Web technologies. The RML-based methodology was introduced alongside the Aldrovandi temporal exhibition case study on which it was tested and refined, with particular attention to balancing the automation of replicable steps and the necessity of case-specific ad hoc customizations.

In particular, we outlined the value of human contribution in the context of data management in a workflow for the RDF formalization of the metadata collections concerning the temporary exhibition's objects and their digital acquisition process. Significant emphasis was placed on the interaction between digital humanists and domain experts and on the formalization of knowledge in machine-readable formats through custom online questionnaires. Potential integrations and adaptations have been outlined, such as extending the range of supported formats, adopting a "Human in the Loop" approach, and introducing a potential solution for an automated, executable, and reproducible workflow.

7. ACKNOWLEDGMENTS

This work was partially funded by Project PE 0000020 CHANGES - CUP B53C22003780006, NRP Mission 4 Component 2 Investment 1.3, funded by the European Union - NextGenerationEU.

REFERENCES

[1] Anderson, Marc, and Karën Fort. 'Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint'. *The International Review of Information Ethics* 31, no. 1 (November 2022). https://doi.org/10.29173/irie477.

[2] Auer, Sören, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann, and David Aumueller. 'Triplify: Light-Weight Linked Data Publication from Relational Databases'. In *Proceedings of the 18th International Conference on World Wide Web - WWW '09*, 621. Madrid, Spain: ACM Press, 2009. https://doi.org/10.1145/1526709.1526793.

[3] Balzani, Roberto, Sebastian Barzaghi, Gabriele Bitelli, Federica Bonifazi, Alice Bordignon, Luca Cipriani, Simona Colitti, and others. 'Saving Temporary Exhibitions in Virtual Environments: The Digital Renaissance of Ulisse Aldrovandi – Acquisition and Digitisation of Cultural Heritage Objects'. *Digital Applications in Archaeology and Cultural Heritage* 32 (2024): e00309. https://doi.org/10.1016/j.daach.2023.e00309.

[4] Barzaghi, Sebastian, Federica Collina, Francesca Fabbri, Federica Giacomini, Alice Bordignon, Roberto Balzani, Gabriele Bitelli, and others. 'Digitisation of Temporary Exhibitions: The Aldrovandi Case'. In *Eurographics Workshop on Graphics and Cultural Heritage*, 181–83. The Eurographics Association, 2023. https://doi.org/10.2312/gch.20231176.

[5] Barzaghi, Sebastian, Ivan Heibi, Arianna Moretti, and Silvio Peroni. 'Developing Application Profiles for Enhancing Data and Workflows in Cultural Heritage Digitisation Processes', 18 April 2024. https://doi.org/10.48550/arXiv.2404.12069.

[6] Bizer, Christian, and Andy Seaborne. 'D2RQ –Treating Non-RDF Databases as Virtual RDF Graphs'. *World Wide Web Internet and Web Information Systems*, 1 January 2005.

[7] Corcho, Oscar, Magnus Eriksson, Krzysztof Kurowski, and Milan Ojsteršek. *EOSC Interoperability Framework: Report from the EOSC Executive Board Working Groups FAIR and Architecture*. Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, 2021. https://www.afs.enea.it/project/madia/Documenti/web/docs/EOSC-Interoperability-Framework-KI0221055ENN.pdf.

[8] Daga, Enrico, Luigi Asprino, Paul Mulholland, and Aldo Gangemi. 'Facade-X: An Opinionated Approach to SPARQL Anything'. In *Studies on the Semantic Web*, edited by Mehwish Alam, Paul Groth, Victor de Boer, Tassilo Pellegrini, Harshvardhan J. Pandit, Elena Montiel, Víctor Rodríguez Doncel, Barbara McGillivray, and Albert Meroño-Peñuela. IOS Press, 2021. https://doi.org/10.3233/SSW210035.

[9] Daga, Enrico, Luca Panziera, Carlos Pedrinaci, Serena Villata, Jeff Z. Pan, and Mauro Dragoni. 'BASIL: A Cloud Platform for Sharing and Reusing SPARQL Queries as Web APIs'. In *CEUR Workshop Proceedings*, Vol. 1486, 2015. https://oro.open.ac.uk/45430/.

[10] Daquino, Marilena, Silvio Peroni, David Shotton, and Arcangelo Massari. 'The OpenCitations Data Model'. *Semantic Web Conf. 12507*, 2020. https://doi.org/10.6084/m9.figshare.3443876.v7.

[11] Dimou, Anastasia, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. 'RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data'. In *Proceedings of the Workshop on Linked Data on the Web Co-*

Located with the 23rd International World Wide Web Conference (WWW 2014), edited by Christian Bizer, Tom Heath, Sören Auer, and Tim Berners-Lee, Vol. 1184. CEUR Workshop Proceedings. Seoul, Korea: CEUR-WS.org, 2014. https://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

[12] Ferri, Paolo, Ivan Heibi, Luca Pareschi, and Silvio Peroni. 'MITAO: A User Friendly and Modular Software for Topic Modelling'. *PuntOorg International Journal* 5, no. 2 (24 October 2020): 135–49. https://doi.org/10.19245/25.05.pij.5.2.3.

[13] Heibi, Ivan, Silvio Peroni, Luca Pareschi, and Paolo Ferri. 'MITAO: A Tool for Enabling Scholars in the Humanities to Use Topic Modelling in Their Studies'. In *AIUCD 2021 - Book of Extended Abstracts*, 175–82, 2021. https://doi.org/10.6092/unibo/amsacta/6712.

[14] Heyvaert, Pieter, Ben De Meester, Anastasia Dimou, and Ruben Verborgh. 'Declarative Rules for Linked Data Generation at Your Fingertips!' In *The Semantic Web: ESWC 2018 Satellite Events*, edited by Aldo Gangemi and others, Vol. 213–217. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-98192-5_40.

[15] Kapsalis, Effie, and Smithsonian Institution. *The Impact of Open Access on Galleries, Libraries, Museums, & Archives*. 1 online resource: color illustrations. vols, 2016. https://purl.fdlp.gov/GPO/gpo158194.

[16] Klieve, Helen, Wendi Beamish, Fiona Bryer, Robyn Rebollo, Heidi Perrett, and Jeroen Van Den Muyzenberg. 'Accessing Practitioner Expertise Through Online Survey Tool LimeSurvey', 2010. http://hdl.handle.net/10072/36611.

[17] Koster, Lukas, and Saskia Woutersen-Windhouwer. 'FAIR Principles for Library, Archive and Museum Collections: A Proposal for Standards for Reusable Collections'. *The Code4Lib Journal*, no. 40 (4 May 2018). https://journal.code4lib.org/articles/13427.

[18] Lefrançois, Maxime, Antoine Zimmermann, and Noorani Bakerally. 'Flexible RDF Generation from RDF and Heterogeneous Data Sources with SPARQL-Generate'. In *Knowledge Engineering and Knowledge Management*, edited by Paolo Ciancarini and others, 10180:131–35. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-58694-6_16.

[19] Massari, Arcangelo, Silvio Peroni, Francesca Tomasi, and Ivan Heibi. 'Representing Provenance and Track Changes of Cultural Heritage Metadata in RDF: A Survey of Existing Approaches', 2023. https://doi.org/10.48550/ARXIV.2305.08477.

[20] Ojsteršek. 'Crosswalk of Most Used Metadata Schemes and Guidelines for Metadata Interoperability'. Zenodo, 5 January 2021. https://doi.org/10.5281/ZENODO.4420115.

[21] Peinado-Santana, Sara, Patricia Hernández-Lamas, Jorge Bernabéu-Larena, Beatriz Cabau-Anchuelo, and José Antonio Martín-Caro. 'Public Works Heritage 3D Model Digitisation, Optimisation and Dissemination with Free and Open-Source Software and Platforms and Low-Cost Tools'. *Sustainability* 13, no. 23 (January 2021). https://doi.org/10.3390/su132313020.

[22] Pescarin, Sofia. 'Museums and Virtual Museums in Europe: Reaching Expectations'. *SCIRES-IT - SCIentific RESearch and Information Technology* 4, no. 1 (30 April 2014): 131–40. https://doi.org/10.2423/i22394303v4n1p131.

[23] Renda, Giulia, Marilena Daquino, and Valentina Presutti. 'Melody: A Platform for Linked Open Data Visualisation and Curated Storytelling', 26 June 2023. http://arxiv.org/abs/2306.14832.

[24] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3, no. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Digitalizzazione e modellazione della *Drammaturgia* di Leone Allacci

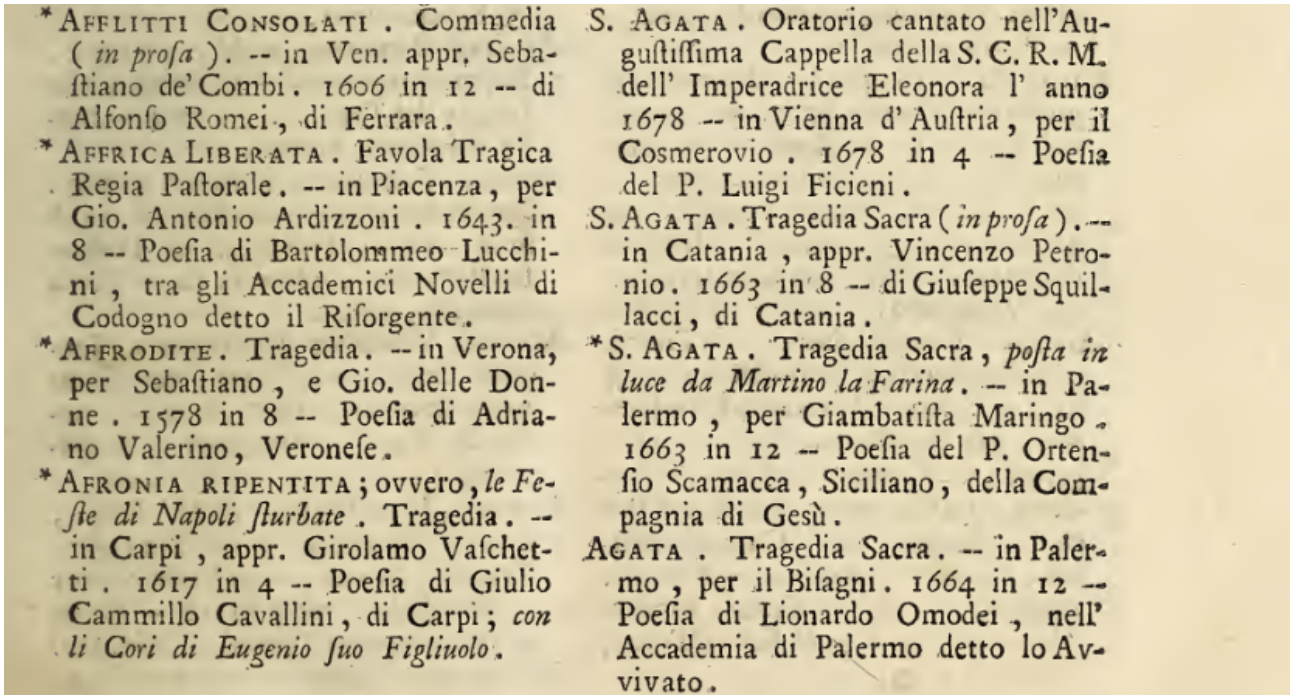
Luca Giovannini [1], Giorgia Gallucci [2]
[1]Università di Potsdam, Germania / Università di Padova, Italia – giovannini@uni-potsdam.de
[2] Università di Padova, Italia – galluccigiorgiag@gmail.com

ABSTRACT
Il contributo presenta il progetto di digitalizzazione e trasformazione in database di uno dei maggiori cataloghi di opere teatrali italiane: la *Drammaturgia* di Leone Allacci (1666), nella sua versione riveduta e ampliata da Giovanni Cendoni, Apostolo Zeno e altri (Venezia, Pasquali, 1755).

PAROLE CHIAVE
Teatro italiano; database; digitalizzazione; modellazione.

1. INTRODUZIONE
Leone Allacci (1586-1669), studioso e teologo greco attivo nell'ambiente romano, svolge un ruolo cardine per la cultura europea seicentesca [9]. Si tratta di una figura centrale negli scambi eruditi ed epistolari del XVII secolo, interlocutore privilegiato di personaggi rilevanti in ambito italiano ed estero come Antonio Magliabechi, Angelico Aprosio, Gabriel Naudé, François Combefis, Giulio Mazzarino e Jean-Baptiste Colbert. Il contributo di Allacci risulta capillare visto che le sue opere vengono pubblicate in diversi ed importanti centri editoriali (Roma, Parigi, Lione, Colonia) e sono numerosissime: sessanta quelle andate a stampa, più di duecento i progetti fermi alla forma manoscritta. La sua copiosa produzione verte su argomenti molteplici che spaziano dalla teologia alla bizantinistica, dall'architettura sacra all'antichità classica fino alla letteratura italiana.
Tra gli esiti di questa pluralità di interessi spicca la *Drammaturgia*, una bibliografia delle opere teatrali in lingua italiana edite ed inedite a cui Allacci, in collaborazione con Aprosio, lavora tra il 1654 e il 1666 – anno della pubblicazione della *princeps* romana, per i tipi del Mascardi. L'opera include inizialmente un elenco alfabetico dei drammi accompagnato da diversi indici: l'indice dei nomi e dei cognomi degli scrittori, delle città d'origine di molti autori, dei soggetti drammatici, delle opere inedite ma menzionate in altri repertori, e delle varianti nei titoli presenti nella *Drammaturgia*. Il testo riscuote un successo inatteso, visto il carattere elencatorio della trattazione, ma non viene ripubblicato in seguito; proprio l'assenza di una ristampa e la difficoltà a trovarne delle copie sono alla base della ripresa settecentesca del lavoro. A metà del XVIII secolo, infatti, un gruppo di letterati della cerchia di Apostolo Zeno, noto poeta e librettista veneziano [1], fornisce una nuova edizione aggiornata, edita a Venezia per il Pasquali (1755). Le informazioni contenute negli indici della *princeps* si inseriscono ora in un elenco unico di drammi, per titoli, ma ricco di dettagli e che lascia spazio anche a riflessioni dei nuovi compilatori su specifiche di carattere editoriale, stilistico, tematico. Questa versione accresciuta offre dunque al lettore un quadro esaustivo del panorama teatrale in Italia tra la fine del Cinquecento e il Settecento.
Da un punto di vista formale, la *Drammaturgia* si presenta come una sorta di database *ante litteram:* soprattutto nella sua veste settecentesca, è un'opera caratterizzata da una struttura coerente e ripetitiva, ordinata secondo il criterio alfabetico e tendenzialmente organica nell'utilizzo dei segni diacritici per distinguere le distinte sezioni informative. Nella maggior parte dei casi, al titolo segue un punto fermo, poi l'indicazione del genere e, tra parentesi, del metro; due lineette precedono le specifiche editoriali (città, stampatore, anno e formato); chiudono, di norma, i riferimenti all'autore e alle sue origini (vd. Fig. 1).
La creazione di un database basato sull'Allacci, che si presenta in questa sede, risponde a una duplice esigenza. Innanzitutto, si intende facilitare l'accesso degli studiosi a un repertorio chiave per le indagini critiche sulla drammaturgia italiana tra Rinascimento e Barocco, finora consultabile solo in cartaceo o in copie digitalizzate non interrogabili, e comunque privo di un'edizione moderna. In secondo luogo, si propone una nuova risorsa per l'esplorazione del patrimonio teatrale italiano dei secoli XVI-XVIII, che integra e sviluppa in termini di modellazione e di fruizione lavori già esistenti ma limitati (come un censimento di testi pastorali [10], anch'esso derivato dalla *Drammaturgia)*. Il *database* intende offrire all'utente la possibilità di orientarsi tra i numerosi *item* dell'opera, accompagnandolo in analisi su autori, generi, centri editoriali, forme metriche e ulteriori dettagli. La validità di simili strumenti di ricerca, peraltro, appare confermata dalla diffusione dell'ormai affermato progetto *Corago* [2], che, seppur limitato ai soli testi operistici, rappresenta a tutti gli effetti lo standard per quanto riguarda la catalogazione di metadati relativi a opere sceniche.

Al di là del certo interesse dell'iniziativa per gli studi umanistici, il progetto costituisce inoltre un *case study* per quanto riguarda lo sviluppo agile di risorse digitali al di fuori di contesti di ricerca più strutturati. L'*Allacci Digitale* si configura infatti come un esperimento di prototipazione e costruzione di un database letterario secondo un approccio che è possibile definire come *low-resources* in termini di costi, tempistiche, e personale richiesto. In questo, l'*Allacci Digitale* differisce da cataloghi frutto di sforzi collettivi come *Corago* e risulta invece affine a strumenti quali il *Database of German-Language One-Act Plays 1740–1850* [4], che è tuttavia parzialmente supportato dalla già esistente infrastruttura *DraCor* [5], e *BIB18* [11], simile nell'impianto ma basato su una bibliografia già digitalizzata.
Una componente cruciale in questo approccio *low-resources* è inoltre l'impiego di modelli linguistici di grandi dimensioni (LLM) a supporto della modellazione, come accaduto nell'elaborazione degli script per l'estrazione delle informazioni. In linea con quanto sostenuto da Andres Karjus, la creazione dell'*Allacci Digitale* dimostra quindi come "tasks previously requiring potentially months of team effort and complex computational pipelines can now be accomplished by an LLM-assisted scholar in a fraction of the time" [7: 1].



*Afflitti Consolati. Commedia (*in proſa*). -- in Ven. appr. Sebaſtiano de' Combi. 1606 in 12 -- di Alfonſo Romei, di Ferrara.
*Affrica Liberata. Favola Tragica Regia Paſtorale. -- in Piacenza, per Gio. Antonio Ardizzoni. 1643. in 8 -- Poeſia di Bartolommeo Lucchini, tra gli Accademici Novelli di Codogno detto il Riſorgente.
*Affrodite. Tragedia. -- in Verona, per Sebaſtiano, e Gio. delle Donne. 1578 in 8 -- Poeſia di Adriano Valerino, Veroneſe.
*Afronia ripentita; ovvero, *le Feſte di Napoli ſturbate*. Tragedia. -- in Carpi, appr. Girolamo Vaſchetti. 1617 in 4 -- Poeſia di Giulio Cammillo Cavallini, di Carpi; *con li Cori di Eugenio ſuo Figliuolo*.
S. Agata. Oratorio cantato nell'Auguſtiſſima Cappella della S. C. R. M. dell' Imperadrice Eleonora l' anno 1678 -- in Vienna d' Auſtria, per il Cosmerovio. 1678 in 4 -- Poeſia del P. Luigi Ficieni.
S. Agata. Tragedia Sacra (*in proſa*). -- in Catania, appr. Vincenzo Petronio. 1663 in 8 -- di Giuſeppe Squillacci, di Catania.
*S. Agata. Tragedia Sacra, *poſta in luce da Martino la Farina*. -- in Palermo, per Giambatiſta Maringo. 1663 in 12 -- Poeſia del P. Ortenſio Scamacca, Siciliano, della Compagnia di Gesù.
Agata. Tragedia Sacra. -- in Palermo, per il Biſagni. 1664 in 12 -- Poeſia di Lionardo Omodei, nell' Accademia di Palermo detto lo Avvivato.

Figura 1. Una pagina della Drammaturgia

2. MODELLAZIONE

La copia digitalizzata dell'opera usata per creare l'*Allacci Digitale* è conservata nel Fondo Cavriani Melzi della Biblioteca storica "Arturo Graf" (Università di Torino)[1] ed è stata sottoposta a trascrizione automatica attraverso il software *Transkribus* [7], utilizzando il modello generalista per testi stampati *Print M1*. Le scansioni, seppur globalmente di buona qualità, presentavano alcune imperfezioni (in termini di stampa e orientamento delle pagine) che hanno reso necessario un lavoro preliminare di definizione manuale delle sezioni di pagina da riconoscere.[2] Il risultato dell'OCR (un file TXT) è stato poi revisionato attraverso diversi round di correzione manuale per emendare errori frequenti di trascrizione; ai fini di una modellazione più agile, si è scelto di limitare le cautele filologiche e di normalizzare, quando ritenuto opportuno, le voci. L'obiettivo, infatti, non è quello di fornire una trascrizione diplomatica della *Drammaturgia*, ma di agevolare la consultazione delle informazioni in essa contenute.
Nella fase successiva sono state sviluppate euristiche per l'estrazione delle informazioni più rilevanti basate su pattern ricorrenti. Le caratteristiche del testo, che si presenta in un formato pressoché tabulare con variabilità limitata, hanno suggerito di usare approccio minimale, basato sulla segmentazione delle voci in corrispondenza di indici significativi, piuttosto che di impiegare tecniche più avanzate basate sull'elaborazione del linguaggio naturale (con annotazione morfosintattica e lemmatizzazione). A tutti gli effetti, si è quindi impiegato un approccio a bassa supervisione (*lightly-*

[1] I file sono accessibili attraverso la nuova Biblioteca Digitale di Unito: https://dl.unito.it/it/. In seguito, è stata reperita una copia di migliore qualità, usata come riferimento per le elaborazioni successive e disponibile nell'*Internet Archive* all'indirizzo https://archive.org/details/drammaturgia00alla/.
[2] Gli autori ringraziano Lara Piva (READ-COOP/Università di Padova) per la consulenza fornita in questa fase.

supervised approach, o LSA), in cui il controllo umano si è limitato "alla definizione [e continuo aggiornamento] delle regole di estrazione definite in base allo specifico ambito di applicazione" [3: 198]. In termini pratici, una serie di script Python[3] ha provvisto a estrarre da ogni voce nove campi significativi: titolo, sottotitolo, autore, genere, metro (prosa/versi), luogo di edizione, editore, anno, formato tipografico (in 4, in 8, in 12, ecc.). Sono state inoltre contrassegnate attraverso parole chiave le opere tradotte e quelle comprendenti un accompagnamento musicale (libretti, melodrammi e forme analoghe).

Terminata questa fase iniziale, che ha permesso una prima esplorazione dei dati (§3), ed esportato il database in formato CSV, si procederà a una sua comprensiva pulizia attraverso lo strumento OpenRefine,[4] utile per correggere gli errori di estrazione ancora presenti[5] ma soprattutto per collegare il maggior numero possibile di entità a piattaforme bibliografiche esterne, come il VIAF o Wikidata, trasformandoli in Linked Open Data. Lo scopo finale è la pubblicazione del database sul sito del progetto,[6] accompagnato da un'opportuna maschera di ricerca e da una serie aggiornata di statistiche testuali.

3. ESPLORAZIONE

Sebbene l'*Allacci Digitale* sia ancora in fase di rifinitura,[7] è già possibile condurre una prima analisi esplorativa dei suoi contenuti, che rivela le potenzialità del *database* come supporto per un'investigazione quantitativa del teatro italiano dal Cinquecento al Settecento. I grafici riportati in Fig. 2, ad esempio, forniscono informazioni significative da un punto di vista temporale, geografico, formale e autoriale. L'interpretazione di questi dati, ad esempio, permette di evidenziare una tendenza al recentismo: le voci catalogate si addensano nella contemporaneità sia dell'Allacci sia dei continuatori del suo progetto, e ciò appare una possibile conseguenza della facilità di reperimento delle informazioni, ma anche dell'interesse dei compilatori per il panorama a loro contemporaneo.

Allo stesso modo, non stupisce il primato di Apostolo Zeno tra gli autori censiti nella *Drammaturgia*, visto il ruolo centrale che assume nella curatela dell'edizione del 1755. È degno di nota il fatto che accanto al suo nome si trovino largamente attestati altri librettisti (Niccolò Minato, Francesco Silvani, Pietro Metastasio, Matteo Noris), a dimostrazione dell'importanza che il teatro per musica riveste nel periodo preso in considerazione. Tale rilievo risulta meno evidente in una classificazione per generi proprio per la variabilità e l'incertezza nella denominazione della produzione operistica. A partire da queste considerazioni preliminari, è già possibile impostare una riflessione sulla ricezione e sulla considerazione della librettistica nello spazio coevo. In maniera analoga, l'indagine spaziale permetterebbe di ragionare non solo su dati già noti, come il dominio della stampa veneziana, ma di mettere in relazione i centri di produzione culturale con altre informazioni fornite dalla *Drammaturgia*.

Gli sviluppi futuri del progetto prevedono sia una rifinitura della trascrizione di base (integrando, ad esempio, le informazioni con i dati ricavati dalla sezione "Aggiunte e correzioni" posta alla fine dell'edizione del 1755) sia un ulteriore arricchimento della base di dati, sviluppando metodi di estrazione per informazioni di accesso meno immediato quali 'librettista', 'luogo di rappresentazione', 'dedicatario', 'occasione', 'provenienza degli autori', ecc. Particolarmente complessa risulta la modellazione delle ristampe, che permetterebbe tuttavia di mappare diacronicamente e diatopicamente il successo e la ricezione di una data opera.

In una prospettiva di interoperabilità, sarebbe inoltre auspicabile collegare i testi operistici catalogati dall'Allacci con il sistema *Corago* e le sue estese informazioni bibliografiche. In ultima battuta, si aspira infine a realizzare un'esplorazione computazionale del database che vada oltre le mere statistiche e si leghi ad attuali temi di ricerca nella storia della letteratura teatrale: una prima ipotesi potrebbe riguardare, ad esempio, il *topic modelling* di titoli e sottotitoli sul modello di [8].

[3] Lo script è disponibile all'indirizzo: https://github.com/allacci-digitale/allacci-digitale.github.io/blob/main/data/modelling-notebook.ipynb.

[4] Vedi https://openrefine.org.

[5] Attualmente la media delle percentuali di mancata estrazione (dovuta all'effettiva assenza del campo nella voce o ad errori di estrazione per via di regex incomplete o insufficienti) è inferiore al 5%; i campi "autore" e "editore" risultano i più problematici da gestire.

[6] Il sito è consultabile all'indirizzo https://allacci-digitale.github.io.

[7] Il database (v1.0.0) è disponibile in formato JSON e CSV sia sul sito web del progetto (all'indirizzo https://allacci-digitale.github.io/database.html, sezione "Downloads") sia su *Zenodo* (https://doi.org/10.5281/zenodo.10972670).

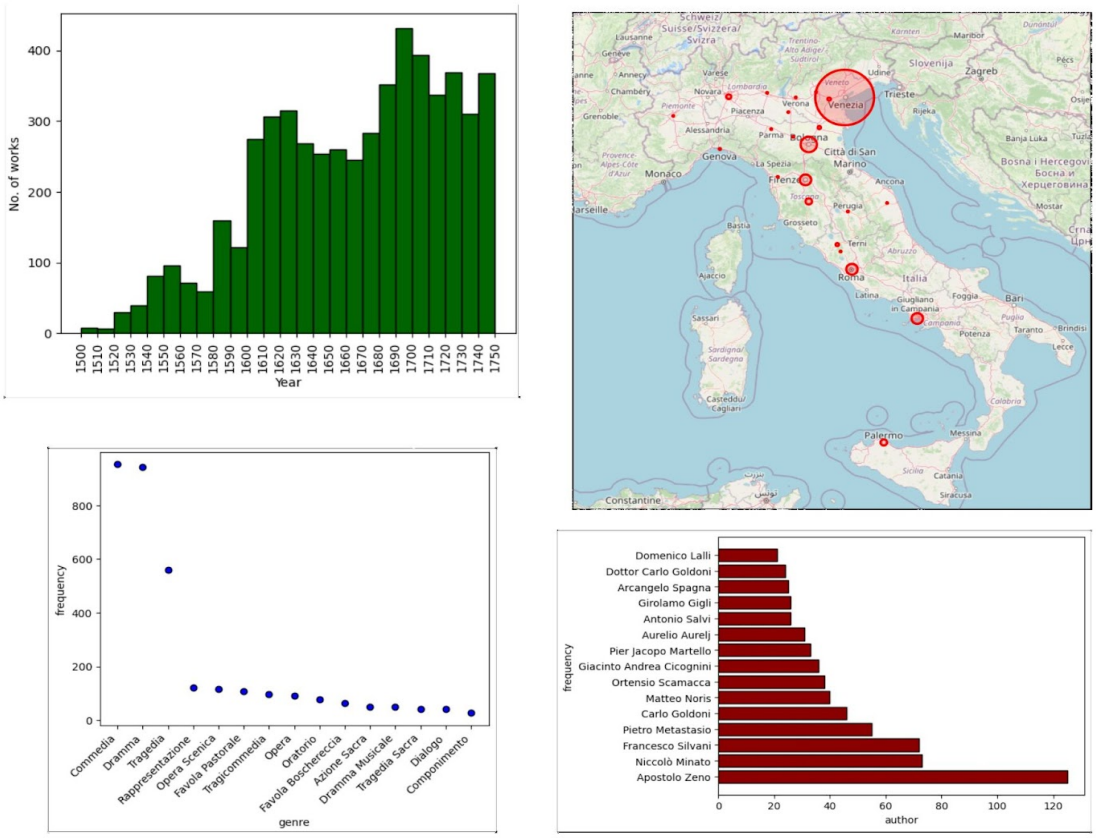


Figura 2. Alcuni esempi di statistiche estratte dalla base di dati. Da sinistra, in senso orario: distribuzione temporale delle opere, luoghi di pubblicazione più frequenti (top 25), autori più frequenti (top 15), generi più frequenti (top 15).

BIBLIOGRAFIA

[1] Bizzarini, Marco. «Zeno, Apostolo». In *Dizionario Biografico degli Italiani*, 2020. https://www.treccani.it/enciclopedia/apostolo-zeno_%28Dizionario-Biografico%29/.

[2] Bonora, Paolo. «Impiego del Web Semantico per lo sviluppo e la consultazione di archivi musicali. Un caso di studio sulla storia e la documentazione del melodramma italiano: l'archivio Corago». Tesi di dottorato, XXXII ciclo, Università di Bologna, 2020. https://doi.org/10.6092/unibo/amsdottorato/9174.

[3] Bonora, Paolo, e Angelo Pompilio. «Estrazione automatica delle caratteristiche del personaggio d'opera attraverso pattern lessico-sintattici.» *Umanistica Digitale* 10 (2021). https://doi.org/10.6092/issn.2532-8816/12426.

[4] Çakir, Dîlan Canan, e Franz Fischer. «Dramatische Metadaten. Die Datenbank deutschsprachiger Einakter 1740–1850». In *DHd2022 Book of Abstracts*. Università of Potsdam, 2022. https://doi.org/10.5281/zenodo.6327977.

[5] Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, e Peer Trilcke. «Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama». In *DH2019 Book of Abstracts*. Università di Utrecht, 2019. https://doi.org/10.5281/zenodo.4284002.

[6] Kahle, Philip, Sebastian Colutto, Günter Hackl, e Günter Mühlberger. «Transkribus - a service platform for transcription, recognition and retrieval of historical documents». In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 4:19–24. Università di Kyoto, 2017. https://doi.org/10.1109/ICDAR.2017.307.

[7] Karjus, Andres. «Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence». *arXiv*, 2023. https://doi.org/10.48550/arXiv.2309.14379.

[8] Moretti, Franco. «Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)». *Critical Inquiry* 36, fasc. 1 (2009): 134–58. https://doi.org/10.1086/606125.

[9] Musti, Domenico. «Allacci, Leone». In *Dizionario Biografico degli Italiani*, 1960. https://www.treccani.it/enciclopedia/leone-allacci_%28Dizionario-Biografico%29/.

[10] Pasqualetto, Giuliano. «Per una bibliografia del teatro pastorale in Italia. Rielaborazione ragionata della Drammaturgia di Leone Allacci», 2017. http://www.giulianopasqualetto.it/files_uploads/testi/boschi_amorosi/allacci_biblio.pdf.

[11] Schöch, Christof. «Curation and Analysis of 'XVIIIe siècle: Bibliographie». In *DHd 2024 Book of Abstracts*. Università di Passavia, 2024. https://doi.org/10.5281/zenodo.10698424.

From Data Complexity to User Simplicity: A Framework for Linked Open Data Reconciliation and Serendipitous Discovery

Marco Grasso[1], Giulia Renda[2], Marilena Daquino[3]
[1] University of Bologna, Italy - marco.grasso7@unibo.it
[2] University of Bologna, Italy - giulia.renda3@unibo.it
[3] University of Bologna, Italy - marilena.daquino2@unibo.it

ABSTRACT
This article introduces a novel software solution to create a Web portal to align Linked Open Data sources and provide user-friendly interfaces for serendipitous discovery. We present the Polifonia Web portal as a motivating scenario and case study to address research problems such as data reconciliation and serving generous interfaces in the music heritage domain.

KEYWORDS
Linked Open Data; Data Reconciliation; Generous interfaces; Music Heritage.

1. INTRODUCTION
Linked Open Data (LOD) offers immense potential for data integration and knowledge discovery in the Cultural Heritage domain [2]. However, a significant challenge remains the alignment of several LOD sources while simultaneously providing user-friendly interfaces (UI) to diverse stakeholders [20].
In this article we present a reusable software solution and user interfaces we developed to address such problems, and we present the Polifonia Web portal as a case study. The Polifonia Web portal[1] addresses such challenges by providing models, methods, and interfaces tailored to (1) bridge connections within the rich and diverse domain of music heritage, and (2) foster engagement of a broad audience interested in serendipitous discovery, rather than experts' tasks only. It is designed to access a registry of music resources available online, using LOD as lingua franca to share and retrieve data. The portal offers tools to perform data aggregation and alignment, both at instance level (i.e., retrieving equivalent entities across datasets) and at ontology level (i.e. offering customizable query strategies to access sources adopting diverse ontologies). Secondly, data populate user-friendly interfaces designed to address informative needs of domain experts as well as to stimulate curiosity in the general audience, which may not have a specific task in mind.
After a concise review of the state of the art of User Interfaces for LOD-based reconciliation and exploration, we outline our methodological approach, highlight the outcomes of the research, describe the components of the proposed framework, and present the case study to validate our approach.

2. STATE OF THE ART
The state of the art in the field of digital data management, discovery, and interface design, particularly in the context of LOD and digital heritage, reveals several gaps and emerging challenges [10]. Existing solutions (such as *Omeka S*, *Semantic MediaWiki*, *Sinopia* and *ResearchSpace*) often fall short in key areas such as user-friendly interface building, provenance management, integration with existing data management workflows (including versioning, backup, and long-term preservation), and the reusability and sustainability of produced assets and software. These limitations are critical, as they impact the quality and sustainability of data management practices [9, 17, 19].
When addressing interfaces for displaying digital cultural heritage, research tends to split into two areas, namely: developing exploratory interfaces for the public focused on enjoyment and serendipity [7], and more complex, hypothesis-driven LOD analytical tools for experts [13]. While the concept of generous interfaces [18] aids in designing public-facing tools, effective mechanisms for LOD exploration and storytelling are notably absent. In recent works [5], 77 Linked Data visualization tools have been surveyed and evaluated. Results show that such tools are not sufficiently equipped for providing full support to users who want to explore and gain knowledge from Linked Data. Some issues regard scalability, the provision of dataset statistics beyond generic counters, and the possibility to combine different visualizations into a user-defined narrative. To the best of our knowledge, no out-of-the-box solutions exist to facilitate the alignment of LOD sources and, at the same time, populate user-friendly interfaces.

[1] https://polifonia.disi.unibo.it/portal

3. APPROACH

We developed a software solution that addresses the above gaps. It facilitates the alignment of LOD sources, offering customizable data reconciliation and data indexing options, as well as the population of user-friendly interfaces including text search features and data mashup techniques to generate descriptive web pages. Our approach to designing such a web solution merges eXtreme Design [11] and Design Thinking [6, 12, 15] methodologies. Initially, the ontology design team develops personas and competency questions with input from Polifonia experts and stakeholders. Competency questions are analysed, both qualitatively and quantitatively [3, 4] to understand data and user journeys, classifying them according to their complexity and the (lack of) task at hand. A competitive analysis of existing LOD-native storytelling solutions, and web applications in cultural heritage was performed, complemented by focus groups with experts and user studies with lay users, which informed our design process, ensuring our solutions are both innovative and user-centric [8, 16]. We summarised insights to outline interaction patterns, selected deployment solutions, and conducted user testing sessions for usability validation and co-design purposes. This streamlined approach [14] ensures our design is not only advanced but also resonates well with both experts and the general public. By incorporating principles of co-design throughout this process, we ensure that our solutions are developed not just for users but with them, fostering a sense of ownership and relevance in the final product [1].

4. RESULTS: THE POLIFONIA WEB PORTAL FRAMEWORK

The Polifonia Web portal serves as the primary gateway to the vast array of musical heritage digital assets curated by project partners. The portal's key features include overall views of categories relevant to the domain, featured highlights to exemplify a user journey, text search functions, and the provision of on-demand insights on entities that are part of the knowledge graph. Fig.1 illustrates the main components and features that characterize the Web portal.

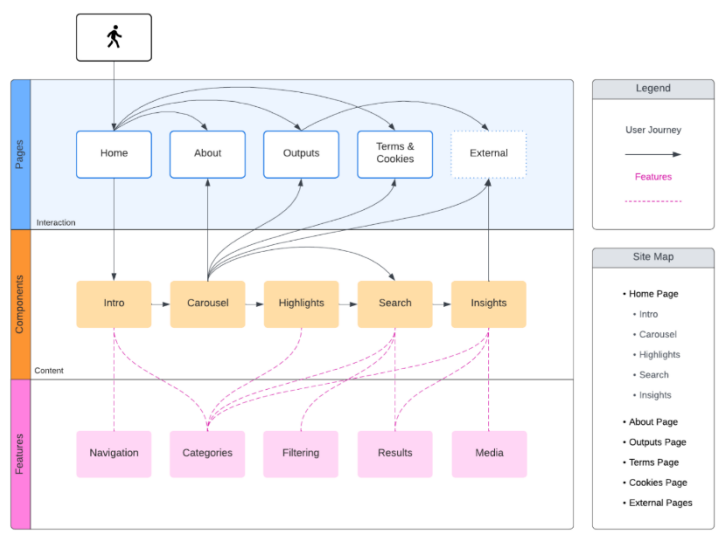


Figure 1. Pages, components, and features of the Web portal

The Web portal is composed of a well-orchestrated array of web pages along with a suite of interactive UI components and actionable features. The **Home page** guides users from a broad understanding of the functionalities available on the Web portal to more specific ones. The **Intro** component shows a quick take home message, and a **Carousel** shows key pages and resources connected to the Web portal (see Fig. 2a).

As the user scrolls, the **Highlights** section offers a list of example dataweb available on the platform (see Fig. 2b). Like a music box playing tunes, this section shows featured content and navigation options, creating a welcoming introduction to the diverse access points to the datasets indexed in the Web portal. Each highlight represents a category (in the Polifonia Web portal we adopted 5 categories, namely: genres, artists, places, music and instruments) which corresponds later in the same homepage to a section with a dedicated text search functionality. Each highlight is accompanied by a sound (which users can toggle on or off) and a colored dot. The color is encoded using the Polifonia palette and it is consistently used in subsequent sections to represent a certain category, e.g. yellow for 'places'. Highlights are displayed in a five-column layout, mirroring a music box, with each column symbolizing one of the five categories mentioned earlier. Notably, highlighted

entities are linked by an associative relationship. Clicking on a highlight takes users directly to the corresponding search section.

Search sections allow users to explore data connected to an entity, performing a lookup search (see Fig. 2c). The text search returns a list of autocompletion suggestions retrieved from an index. Each **index** includes the list of entities of the same type from different datasets, integrating cross-dataset alignments to minimize duplicates in the results. Results of the search represent relations between the search entity and other entities included in the ingested datasets. Relations between the topic and other entities are listed below the search bar, and users can filter by type of relation, category of connected entities, and source of information.

Lastly, the **Insights card** allows users to expand search results and delve into details of linked entities (see Fig. 2d). A card mainly includes texts, links to external resources, multimedia, and associations between the entity at hand and other entities belonging to any of the ingested or linked datasets (e.g. Wikidata, DBpedia, Discogs).

Throughout the user journey, the intentional arrangement of components on the page ensures a seamless transition from a general overview to more specific areas of interest and searches, reflecting the original idea of creating generous interfaces. The user journey is meticulously designed to not only cater to diverse needs but also foster serendipitous discoveries by guiding users step by step through our platform's offerings and revealing information on demand. A few use cases have been designed to exemplify the potential of such an approach in real-world scenarios, namely: serendipitous discovery of multimedia connected to an artist (targeted to lay users), interconnected archives (targeted to CH curators), and music tourism guided by the influence of a music genre over nearby places (again targeted to lay users).

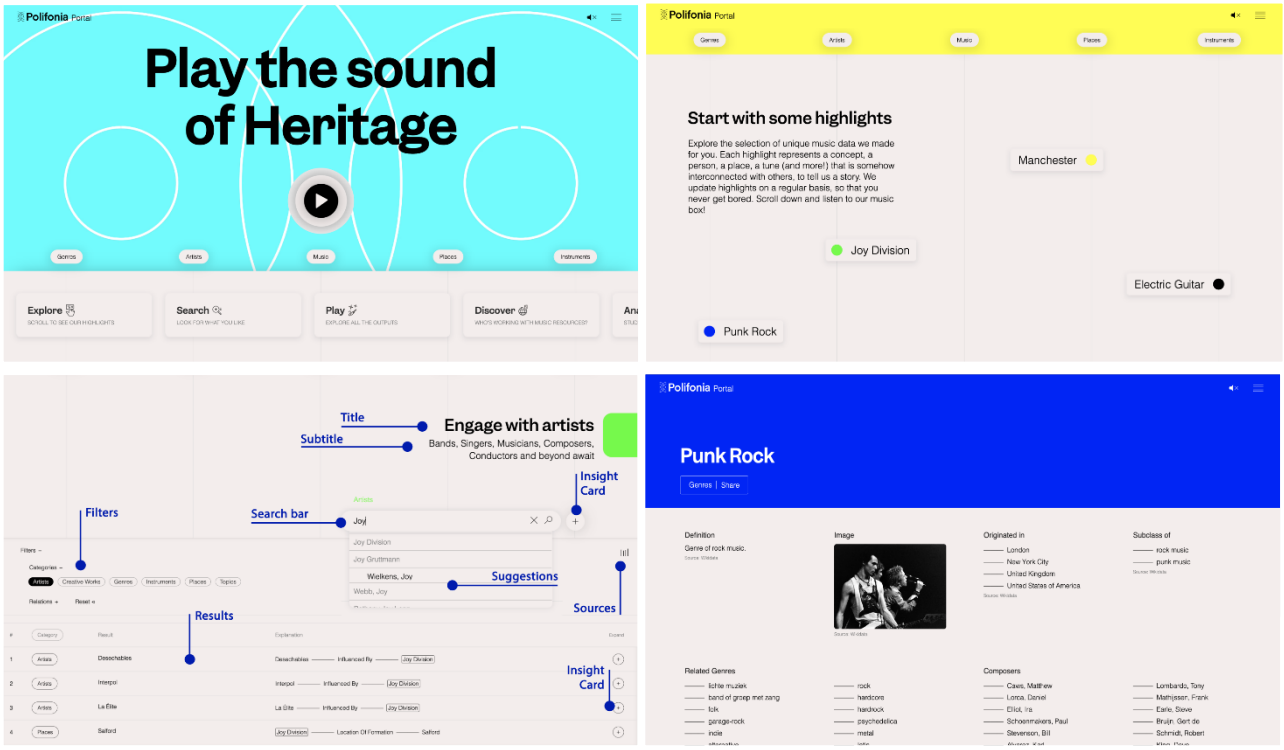


Figure 2. The homepage of the Polifonia Web portal: Intro, Highlights, Search, Insights Card

The Web portal's architecture is designed on top of a *Flask*[2] backend and a *React*-native[3] frontend, which communicate through **RESTful APIs**. The portal uses *Sonic*[4] for efficient data indexing, ensuring a responsive user experience (see Fig. 3). The source code of the application is open source, and it is available on GitHub[5].

The portal is (optionally) complemented by a *Blazegraph*[6] triplestore, which plays a crucial role in storing and managing post-processed data. The data ingestion process is aimed at supporting the alignment and it begins with the extraction of data from indexed sources, as detailed in customizable **configuration files**. Such files include the SPARQL queries

[2] https://flask.palletsprojects.com/en/3.0.x/
[3] https://react.dev/
[4] https://github.com/valeriansaliou/sonic
[5] https://github.com/polifonia-project/portal
[6] https://blazegraph.com/

required to retrieve data from each dataset for each category (which are also customizable). For instance, Polifonia configuration files include SPARQL queries to retrieve all <artists>, <places>, <instruments>, <music works>, <genres> (i.e., the categories) from data sources like Wikidata, Dbpedia, etc. Once extracted, data undergoes the **reconciliation** process. This process aims to harmonize and integrate information about entities from diverse sources, contributing to solving the challenging issue of entity duplication across datasets. For example, if two different sources provide information about the same musical entity, the portal's reconciliation system will perform *sameAs* link detection, perform rule-based inference (e.g. looking for transitive links) and merge data in case of a positive match, therefore presenting the user with a unified URI and data view. The reconciliation process unfolds in several stages, starting with the creation of named graphs for each extracted URI, followed by searches for equivalence statements across Polifonia and third-party datasets like *Wikidata* and *DBpedia*. Relations are subsequently expanded leveraging direct and transitive properties and a careful lookup to identify entities appearing in multiple graphs is performed. The outcome is a **linkset** where newly minted Polifonia URIs are linked to URIs belonging to indexed data sources, which are then used to populate the indexes behind text searches.

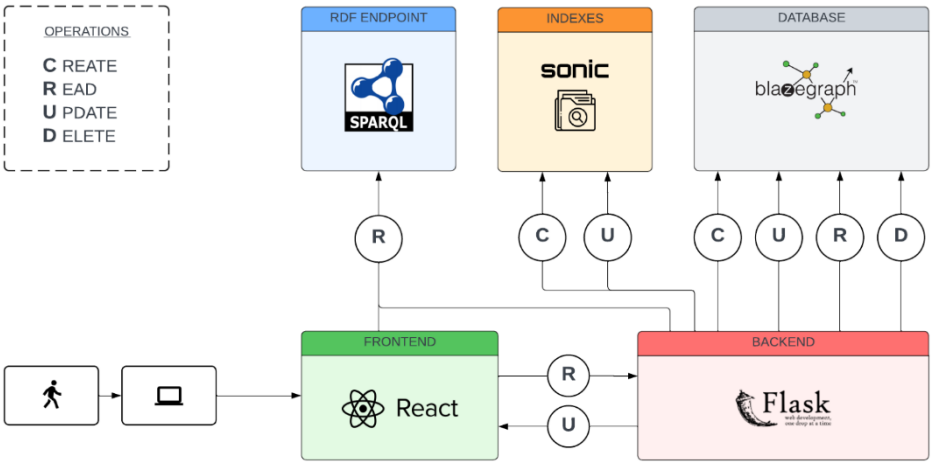


Figure 3. Overview of the technology stack of the Web portal

Customization and reusability are central to the Polifonia Web portal, ensuring it can easily adapt to various datasets, ontologies, and evolving UI/UX needs. This flexibility is achieved through five configuration files, enabling tailored data source access and UI component adjustments. Key customizable features include the ability to (1) add new datasets with comprehensive metadata; (2) define homepage highlights, (3) categories, (4) indexes for populating search sections, and (5) insights templates. Additionally, users can tailor the introductory section with the carousel configuration, specifying titles, descriptions, and interactive elements for each box. The most dynamic feature is the insights cards configuration, which allows for the integration of various content blocks such as text, multimedia, semantic relations, and web links. These blocks are adjustable in size, title, and description, with SPARQL queries fetching relevant content connections, offering a rich, user-centric interface.

The setup allows for the **alignment** of various datasets, making the portal not only versatile in accommodating data external to Polifonia but also **scalable** to efficiently handle new datasets and **adapt** to evolving UI requirements.

5. CONCLUSIONS

We have presented a prototypical web framework for data alignment and population of web pages for serendipitous discovery. Its potential impact and efficacy have been validated through multiple user studies and tests involving participants with diverse backgrounds. In total we performed four sessions divided into two phases. The initial phase was performed prior to development and consisted of one session with lay users[7,8] and one with experts[9]. The second series of

[7] Online form: https://docs.google.com/forms/d/1sBYhkzamGAzJ6IOLeZdQkuTXcIFSsxDpYCJDEaiVgyg/viewform?edit_requested=true

[8] Results: https://github.com/polifonia-project/web_portal/blob/main/questionnaires/survey_march2022.csv

[9] Report: https://github.com/polifonia-project/web_portal/blob/main/questionnaires/Polifonia%20Web%20portal_%20Expert%20Usability%20Review.pdf

user tests was carried out again with general users[10,11] to better frame co-design aspects, and with experts[12,13] to test the final prototype. Feedback from these studies guided iterative improvements, ensuring that the portal not only met the technical requirements for effective LOD management but also resonated with users in terms of ease of use, navigation, and overall experience. Moreover, the tests conducted went beyond mere usability validation and delved into the practical utility of the portal in real-world scenarios. Future works will focus on assessing unresolved data quality issues in collaboration with new stakeholders and evaluating the prototype on new pilot datasets.

6. ACKNOWLEDGMENTS

This work is supported by a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746 (Polifonia: a digital harmoniser for musical heritage knowledge, H2020-SC6-TRANSFORMATIONS).

REFERENCES

[1] Bellucci, Andrea, Giulio Jacucci, Veera Kotkavuori, Bariş Serim, Imtiaj Ahmed, and Salu Ylirisku. 'Extreme Co-Design: Prototyping with and by the User for Appropriation of Web-Connected Tags'. In *End-User Development*, edited by Paloma Díaz, Volkmar Pipek, Carmelo Ardito, Carlos Jensen, Ignacio Aedo, and Alexander Boden, 109–24. Cham: Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-18425-8_8.

[2] Bikakis, Antonis, Eero Hyvönen, Stéphane Jean, Béatrice Markhoff, and Alessandro Mosca. 'Special Issue on Semantic Web for Cultural Heritage'. *Semantic Web 12* 2 (2021): 163–67.

[3] Brown, Tim. 'Design Thinking'. *Harvard Business Review* 86, no. 6 (2008): 84–92.

[4] Chasanidou, Dimitra, Andrea Alessandro Gasparini, and Eunji Lee. 'Design Thinking Methods and Tools for Innovation'. In *Design, User Experience, and Usability: Design Discourse*, edited by Marcus Aaron, 12–23. Cham: Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-20886-2_2.

[5] Desimoni, Federico, and Laura Po. 'Empirical Evaluation of Linked Data Visualization Tools'. *Future Generation Computer Systems* 112 (1 November 2020): 258–82. https://doi.org/10.1016/j.future.2020.05.038.

[6] Dorst, Kees. 'The Core of "Design Thinking" and Its Application'. *Design Studies* 32, no. 6 (1 November 2011): 521–32. https://doi.org/10.1016/j.destud.2011.07.006.

[7] Hyvönen, Eero. 'Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery'. Edited by Pascal Hitzler and Krzysztof Janowicz. *Semantic Web* 11, no. 1 (2020): 187–93. https://doi.org/10.3233/SW-190386.

[8] Liedtka, Jeanne. 'Evaluating the Impact of Design Thinking in Action'. *Academy of Management Proceedings* 1 (August 2017). https://journals.aom.org/doi/abs/10.5465/AMBPP.2017.177.

[9] Mayr, Eva, Paolo Federico, Silvia Miksch, Günther Schreder, Michael Smuc, and Florian Windhager. 'Visualization of Cultural Heritage Data for Casual Users'. In *IEEE VIS Workshop on Visualization for the Digital Humanities*, Vol. 1, 2016.

[10] Po, Laura, Nikos Bikakis, Federico Desimoni, and George Papastefanatos. 'Linked Data Visualization: Techniques, Tools, and Big Data'. In *Synthesis Lectures on the Semantic Web: Theory and Technology*, edited by Ying Ding and Paul Groth, 10:XIV–143. Cham: Springer, 2020. https://doi.org/10.2200/S00967ED1V01Y201911WBE019.

[11] Presutti, Valentina, Enrico Daga, Aldo Gangemi, and Eva Blomqvist. 'EXtreme Design with Content Ontology Design Patterns'. In *Proceedings of Workshop on Ontology Patterns*, edited by Eva Blomqvist, Kurt Sandkuhl, Scharffe, Francois, and Svatek, Vojtek. CEUR Workshop Proceedings. CEUR-WS.org, 2009.

[12] Ramaprasad, Arkalgud, and Thant Syn. 'Design Thinking and Evaluation Using an Ontology'. In *Design Science: Perspectives from Europe*, edited by Markus Helfert and Brian Donnellan, Vol. 447. Communications in Computer and Information Science. Cham: Springer, 2013. https://doi.org/10.1007/978-3-319-13936-4_6.

[13] Renda, Giulia, Marilena Daquino, and Valentina Presutti. 'Melody: A Platform for Linked Open Data Visualisation and Curated Storytelling', 1–8. New York: Association for Computing Machinery, 2023.

[14] Renda, Giulia, Marco Grasso, and Marilena Daquino. 'From Ontology Design to User-Centred Interfaces for Music Heritage'. In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, edited by Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 168-172, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[15] Rowe, Peter G. *Design Thinking*. Cambridge (Massachusetts), London: MIT press, 1991.

[10] Online form: https://docs.google.com/forms/d/e/1FAIpQLSdHO92UByme8fwgWUwAMZV9HNzMrNAb_XthgRN2xSjKzS3MGg/viewform
[11] Results: https://zenodo.org/records/10444375
[12] Online form: https://forms.gle/y8Teo4LNrkvsgafPA
[13] Results: http://doi.org/10.5281/zenodo.10444458

[16] Smuc, Michael, Eva Mayr, and Hanna Risku. 'Is Your User Hunting or Gathering Insights? Identifying Insight Drivers across Domains'. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization, BELIV '10*, 49–54. New York: Association for Computing Machinery, 2010. https://doi.org/10.1145/2110192.2110200.

[17] Villa, Robert, Paul D. Clough, Mark M. Hall, and Sophie A. Rutter. 'Search or Browse? Casual Information Access to a Cultural Heritage Collection'. In *EuroHCIR*, 19–22. Citeseer, 2013.

[18] Whitelaw, Mitchell. 'Generous Interfaces for Digital Cultural Collections'. *Digital Humanities Quarterly 009* 1 (21 May 2015).

[19] Wilson, Max L., and Elsweiler David. 'Casual-Leisure Searching: The Exploratory Search Scenarios That Break Our Current Models'. In *Proceedings of the Fourth Workshop on Human-Computer Interaction and Information Retrieval*, 28–31. New Brunswick, NJ, USA, 2010.

[20] Zahidi, Zaihasriah, Yan Peng Lim, and Peter Charles Woods. 'Understanding the User Experience (UX) Factors That Influence User Satisfaction in Digital Culture Heritage Online Collections for Non-Expert Users'. In *2014 Science and Information Conference*, 57–63. London, 2014. https://doi.org/10.1109/SAI.2014.6918172.

Giuseppe Chiarini: un'opera inedita

Elena Almangano[1], Mirko Castaldi[2], Eleonora De Longis[3], Daniele Pasqualetti[4]
[1]Istituto Italiano di Studi Germanici, Italia - almangano@studigermanici.it
[2]Università Roma Tre, Società Geografica Italiana, Italia - mirko.castaldi@uniroma3.it
[3]Istituto Italiano di Studi Germanici, Italia - delongis@studigermanici.it
[4]Università Roma Tre, Società Geografica Italiana, Italia - daniele.pasqualetti@uniroma3.it

ABSTRACT

Nel 2022 l'Istituto Italiano di Studi Germanici (IISG) ha concluso il progetto di descrizione, ordinamento e inventariazione del proprio archivio storico. L'operazione è stata di grande rilievo in quanto tutto l'inventario è stato messo a disposizione tramite la piattaforma Archiui e condiviso con il Sistema Archivistico Nazionale; questa operazione ha permesso di rendere accessibile e fruibile tutto il complesso documentale. Il riordino ha fatto emergere una parte dell'archivio personale di Giuseppe Chiarini contenente scritti, materiale di studio e, soprattutto, gli appunti per un'antologia inedita dei racconti di viaggio in Italia pubblicati da vari scrittori di area tedesca a lui contemporanei. La scoperta ha già ispirato la realizzazione di un podcast per l'edizione 2023 di Archivissima; con questa prosecuzione di progetto si vuole realizzare una mostra digitale per ampliare la visibilità per un'opera inedita di una figura fondamentale nella storia della cultura italiana. Approfondendo tale tematica, si intende presentare un poster riguardante l'utilizzo di una piattaforma digitale sostenibile per la divulgazione e la fruizione dei fondi archivistici recentemente recuperati con l'obiettivo di incentivarne la conoscenza e lo studio.

PAROLE CHIAVE

Best practice; Digital humanities; Public humanities, Mapping.

1. INTRODUZIONE

Sappiamo bene che gli archivi ci riservano spesso grandi scoperte: è questo stato anche il caso dell'esperienza dell'IISG. Il recente riordinamento dei documenti dell'archivio storico ha fatto emergere gli scritti inediti di Giuseppe Chiarini per la realizzazione di un'antologia odeporica di *excerpta* dai diari di viaggio in Italia di autori tedeschi tra fine '700 e inizio '800. I documenti erano compresi tra quelli del nipote, Paolo Chiarini, donati all'Istituto di cui è stato direttore dal 1968 al 2006. Giuseppe Chiarini (1833-1908) è stato saggista e critico letterario, collaboratore di varie riviste di cultura, biografo di Carducci, Leopardi e Foscolo. Dopo l'Unità ricoprì varie cariche pubbliche e collaborò a stretto contatto con i primi responsabili del ministero dell'Istruzione. Preside di liceo prima a Livorno e poi a Roma, nel 1885 ottenne la cattedra di letterature moderne alla Facoltà di lettere dell'Università di Roma. Pur impegnato nel suo ruolo istituzionale non tralasciò la ricerca erudita e un certo impegno sociale come democratico anticlericale; fra l'altro diresse il periodico «La Domenica del Fracassa» dal 1884 al 1886.

Il fondo è composto prevalentemente di corrispondenza e carte di studio. Le lettere hanno subito in passato un primo e molto impreciso ordinamento, forse a opera dello stesso Chiarini o di un erede. I documenti sono stati semplicemente divisi in ordine alfabetico in base al mittente e raccolti in una camicia (a volte costituita da una lettera usata per contenere tutte le altre). Sui singoli pezzi sempre una mano non identificata ha apposto (sovente a lapis blu) il cognome del mittente (come nel caso in cui la consistenza del fascicolo nominale sia rappresentata da un singolo documento.). In alcuni casi la camicia riporta un elenco di corrispondenti che non sempre rispecchia quelli effettivamente contenuti; anzi a volte vengono menzionati mittenti o destinatari a cui non corrisponde alcun documento. Solo per Pascoli troviamo una camicia nuova (foglio A4 ripiegato), con grafia a noi contemporanea, ma non è stato possibile chiarire chi abbia fatto questa operazione.

Per quanto riguarda i materiali di studio, i singoli fascicoli sono composti da camicie con l'indicazione degli autori della letteratura tedesca di cui si vuole trattare e al loro interno i singoli sotto-fascicoli contengono fogli numerati consecutivamente con *excerpta* delle loro opere. Data la metodica organizzazione dei documenti operata dallo stesso Chiarini, con una originaria divisione per autore in ordine alfabetico-cronologico corredato anche da una bozza di indice nei nomi e dei luoghi al fascicolo 16 si può avanzare l'ipotesi che Chiarini stesse redigendo un'antologia.

L'inventario di tutti i documenti è stato già messo a disposizione tramite la piattaforma digitale Archiui (https://studigermanici.archiui.com/oggetti/2102-giuseppe-chiarini): inoltre, la stessa piattaforma prevede la possibilità di inserire e metadatare alcune scansioni dei documenti.

Parte del materiale è stato utilizzato per realizzare un podcast per l'edizione 2023 di Archivissima, sul tema *Carnet de Voyage*, il cui link è disponibile in sitografia[1].

2. QUALE CULTURA?

Le pagine di Chiarini fanno emergere chiaramente l'impressione suscitata dall'Italia agli scrittori tedeschi che compivano il famoso viaggio nel nostro paese per completare i propri studi. Le città che vengono maggiormente citate sono Milano, Venezia, Firenze, Roma, Napoli, alcuni paesaggi rupestri dell'Umbria e anche città meno frequentate della Basilicata, Puglia e Sicilia, arrivando a dare una panoramica quasi completa dell'Italia di fine '700. A questo proposito, è molto interessante la mappa stilizzata tracciata da Chiarini dei viaggi di Schopenhauer (vd. Fig. 1) che di fatto esemplifica gli itinerari da lui seguiti e le principali mete toccate dagli studiosi stranieri in Italia. È probabile che mappe del genere fossero previste anche per altri autori, almeno per quelli più noti, ma purtroppo non abbiamo documenti che lo attestino.

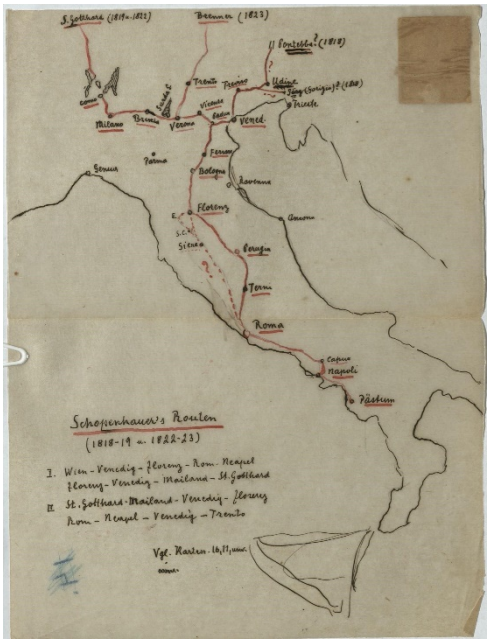


Figura 1. Mappa del viaggio di Schopenhauer

Ogni camicia, che cominceremo a chiamare capitolo, è composta da una breve introduzione biografica dell'autore in questione, stilata da Chiarini sfruttando la sua vasta conoscenza delle opere. Seguono alcuni fogli con antologie di testi in tedesco riguardanti la città o il luogo visitato di cui l'autore ha voluto dare traccia. Questi testi a volte sono trascritti, altri sono ritagliati da fonte sconosciuta e incollati al foglio di scrittura. I testi sono citati in lingua originale, talvolta con note che specificano il senso di alcune parole; questa pratica fa supporre che Chiarini volesse fornire la traduzione italiana.

Lo studioso aveva ipotizzato di suddividere gli autori più corposi con sottocapitoli per affrontare le varie opere o i diversi pensieri che gli erano sorti dall'esperienza di viaggio in Italia. Riprendendo l'esempio di Schopenhauer, il capitolo a lui dedicato viene suddiviso in 3 sottosezioni corrispondenti a: *Il mondo come rappresentazione, Giacomo Leopardi e Impressioni d'Italia*[2].

Molto interessante anche il fatto che tutti i fogli siano numerati, probabilmente dallo stesso autore, e che presentino già dei rimaneggiamenti; una prima stesura conteggiava 641 fogli tutti segnati con lapis blu, tuttavia, poi alcuni fogli vengono eliminati e rinumerati con un lapis rosso proprio ad indicare cambiamenti redazionali di posizione o interi rifacimenti delle sezioni.

A partire dal quarto fino al quindicesimo, quasi tutti i capitoli contengono una camicia di fogli di brutta e appunti di lavoro sul capitolo stesso: anche questi sono molto interessanti perché ci aiutano a entrare nel flusso di lavoro, dal punto di vista sia pratico sia concettuale di Chiarini, individuando indizi su possibili criteri di scelta o sui testi valutati e poi scartati. Sicuramente lo studioso aveva condiviso con qualche amico e corrispondente le sue intenzioni che potrebbero emergere da un'analisi approfondita del carteggio.

[1] Istituto Italiano di Studi germanici. «Impressioni d'Italia: un'antologia a cura di Giuseppe Chiarini». https://www.archivissima.it/2023/oggetti/3101-impressioni-d-italia-unantologia-a-cura-di-giuseppe-chiarini.

[2] Istituto Italiano di Studi germanici. «Giuseppe Chiarini». Istituto Italiano di Studi Germanici. Archivio storico. https://studigermanici.archiui.com/oggetti/2102-giuseppe-chiarini.

Come accennato, questi documenti sono già stati descritti tramite le categorie archivistiche e l'inventario dettagliato è disponibile sulla piattaforma Archiui. Le digitalizzazioni saranno descritte seguendo le indicazioni messe a punto dall'Istituto centrale per gli archivi e tramite il set di metadati descrittivi METS-SAN[3].

3. LA MOSTRA VIRTUALE: UN PROGETTO DI *DIGITAL* E *PUBLIC HUMANITIES*

Un'opera inedita ritrovata a più di cento anni dalla sua progettazione pone sempre un interrogativo rispetto alle modalità di una sua pubblicazione. L'operazione di Chiarini è sicuramente di valore e, ancora oggi, assolutamente originale e la sua messa a disposizione potrebbe sicuramente stimolare un nuovo interesse e nuovi studi a riguardo.

All'inizio di ogni progetto di *Digital* e *Public Humanities* è sempre bene definire un preciso piano di lavoro e suddividerlo in piccoli step con precise indicazioni e task raggiungibili, nonché calibrati in base alle risorse disponibili [1; 2; 3; 5]. Prima di realizzare un'edizione digitale di questo materiale - fase alla quale si pensa anche tramite il progetto parallelo DiScEPT -, obiettivo primario del progetto è la messa a disposizione del patrimonio documentale e la fruizione dello stesso affinché si instaurino nuovi filoni di indagine e possa venire promosso un progetto culturale basato su una documentazione di cui si ignorava prima l'esistenza. Per fare questo intendiamo curare con particolare attenzione i metadati che saranno applicati al singolo oggetto digitale per raggiungere gli obiettivi di interoperabilità, riuso e sostenibilità.

A tale scopo, abbiamo individuato nella formula della mostra virtuale la modalità migliore per rendere fruibile a un pubblico vasto questi documenti e darne una prima classificazione puntuale, tenendo presenti le più recenti acquisizioni nell'ambito delle *Digital* e delle *Public Humanities*.

In proposito abbiamo delineato una sorta di *road map* che ci consenta di definire bene i passi da compiere, applicando le buone pratiche condivise dalla comunità scientifica per rendere il progetto funzionale e sostenibile in ogni sua fase.

Passiamo, quindi, a definire le caratteristiche di ogni fase progettuale, posto che il materiale individuato sia senza dubbio di pregio e che gli originali siano tutti conservati presso l'Archivio storico dell'Istituto Italiano di Studi Germanici. Prendendo come riferimento gli standard previsti dal *Piano nazionale di digitalizzazione del patrimonio culturale* [4], la digitalizzazione, in parte già cominciata, viene effettuata usando uno scanner planetario messo a disposizione dallo stesso ente conservatore. Lo scanner permette di ottenere l'immagine master in TIFF 6.0 non compresso a 600 dpi ottici e con profondità di colore a 24 bit RGB. Per ogni singola immagine in TIFF verrà poi estratta una copia a bassa risoluzione destinata alla piattaforma digitale in JPEG, 150 dpi ottici e profondità di colore di 24 bit RGB.

Lo scanner è dotato di un software di cattura dell'immagine che consente di ottenere per ogni scatto i metadati tecnici relativi, che verranno poi trasportati nella descrizione di ogni immagine all'interno della mostra. Le immagini saranno catturate tramite il riconoscimento ottico dei caratteri (OCR) per poterne poi sfruttare le potenzialità. Il file master verrà temporaneamente conservato sia su *hard disk* esterno sia sul *server* di Istituto per poi essere riversato sul cloud messo a disposizione tramite la Strategia Cloud Italia come da Data Management Plan del Piano nazionale di digitalizzazione del patrimonio culturale.

Si cercherà di evitare quanto più possibile la manipolazione dell'immagine, se questa fosse necessaria, ne sarà dato conto all'interno della scheda descrittiva dell'immagine.

Durante l'elaborazione del progetto e visto anche il materiale geografico, abbiamo ritenuto determinante la collaborazione della Società Geografica Italiana, nello specifico dell'*expertise* dei ricercatori Mirko Castaldi e Daniele Pasqualetti che si stanno occupando dell'elaborazione geo-cartografica dei dati contenuti nei documenti inventariati (un esempio vd. Fig. 2). In particolare, grazie all'utilizzo degli strumenti digitali, verrà costruito un geo-database in ambiente Gis contenente le informazioni più interessanti presenti nel fondo, come: tragitti percorsi, mete toccate e brani selezionati dei vari autori da parte di Chiarini (vd. Fig. 3). Inoltre, verranno realizzati alcuni prodotti digitali in grado di valorizzare il fondo, ad esempio alcune *Storymaps* che permetteranno la navigazione dei documenti attraverso un'esperienza di geo-narrazione interattiva.

In seguito all'analisi della letteratura e alla valutazione di altre esperienze simili, si è ritenuto di pubblicare la mostra tramite la piattaforma *opensource Omeka*[4], realizzata appositamente per la creazione di mostre digitali (un esempio vd. Fig. 4)[5]. Essa risulta particolarmente adatta in quanto permette di metadatare le immagini con il linguaggio *Dublin core* e consente di ampliare la gamma di informazioni che possiamo mettere a disposizione compresa anche la possibilità di inserire la trascrizione del testo presente in ogni immagine. Siamo consapevoli che questa piattaforma permetta una minima descrizione dei dati, ma il suo impiego sarà utile per testare la fattibilità dell'inserimento dei metadati per poi trasporlo su

[3] Istituto centrale per gli archivi – ICAR. «METS-SAN». https://icar.cultura.gov.it/standard/standard-san/mets-san.
[4] Omeka. «Show case». Omeka classic. https://omeka.org/classic/showcase/.
[5] Istituto Italiano di Studi germanici, Società Geografica Italiana, Università Roma Tre. «Giuseppe Chiarini. Un germanista dell'800». https://gchiariniiisg.omeka.net/items/browse.

Archiui. Questo sarà possibile grazie alla digitalizzazione in OCR e l'uso di *Traskribus* combinato con ChatGPT: un utilizzo innovativo dell'AI generativa che ha già trovato applicazione in altre esperienze simili.

Figura 2. Mappa del viaggio di Schopenhauer sovrapposta alla carta geografica fisica dell'Italia

Figura 3. Geolocalizzazione di tutti i toponimi citati da Chiarini nell'indice dei luoghi

Figura 4. Pagina di esempio della mostra virtuale con Omeka

Le immagini saranno rilasciate con licenza Creative Commons CC BY-NC 4.0 e la mostra virtuale sarà fruibile tramite l'*hosting* dell'ente conservatore e il suo inventario digitale.
Siamo profondamente convinti dell'importanza del recupero del fondo archivistico e della sua diffusione presso la comunità degli studiosi e del pubblico interessato.

BIBLIOGRAFIA

[1] Allegrezza, Stefano, (a cura di). *La digitalizzazione del patrimonio culturale. Linee guida, standard, esperienze*. Collana di studi archivistici. Torre del Lago Puccini: Civita Editoriale, 2021.

[2] Falchetta, Piero. «Guida breve alla digitalizzazione in biblioteca». *Biblioteche oggi* 9 (2000): 52–67.

[3] IFLA. *Linee guida per pianificare la digitalizzazione di collezioni di libri rari e manoscritti*. Tradotto da Gruppo di lavoro della Biblioteca digitale BEIC. IFLA, 2015.

[4] Istituto centrale per la digitalizzazione del patrimonio culturale – Digital Library, (a cura di). *Piano nazionale di digitalizzazione del patrimonio culturale*. V. 1.0., 2022. https://docs.italia.it/italia/icdp/icdp-pnd-digitalizzazione-docs/it/v1.0-giugno-2022/index.html.

[5] Tomasi, Francesca. *Organizzare la conoscenza: digital humanities e web semantico: un percorso tra archivi, biblioteche e musei*. Vol. 39. Biblioteconomia e scienza dell'informazione. Milano: Editrice Bibliografica, 2022.

HERITRACE: Tracing Evolution and Bridging Data for Streamlined Curatorial Work in the GLAM Domain

Arcangelo Massari[1], Silvio Peroni[2]
[1] Digital Humanities Advanced Research Centre(/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Italy - arcangelo.massari@unibo.it
[2] Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Italy - silvio.peroni@unibo.it

ABSTRACT
HERITRACE is a semantic data management system tailored for the GLAM sector. It is engineered to streamline data curation for non-technical users while also offering an efficient administrative interface for technical staff. The paper compares HERITRACE with other established platforms such as OmekaS, Semantic MediaWiki, Research Space, and CLEF, emphasizing its advantages in user friendliness, provenance management, change tracking, customization capabilities, and data integration. The system leverages SHACL for data modeling and employs the OpenCitations Data Model (OCDM) for provenance and change tracking. Future developments include the integration of a robust authentication system and the expansion of data compatibility via the RDF Mapping Language (RML), enhancing HERITRACE's utility in digital heritage management.

KEYWORDS
Data Management System; Data Curation; Provenance; Change Tracking; Semantic Web Technologies.

1. INTRODUCTION
In this paper, we introduce HERITRACE (Heritage Enhanced Repository Interface for Tracing, Research, Archival Curation, and Engagement), a novel semantic data management system which addresses the increasing complexities faced by cultural heritage institutions including galleries, libraries, archives, and museums (GLAM). This system has been developed to support the digital landscape of curating metadata in GLAM institutions. Traditionally, GLAM experts have relied on their interpretative skills and domain knowledge to curate metadata. However, the digitization of cultural heritage data has introduced new challenges, including the representation of data in various machine-readable formats and their preservation in heterogeneous databases. This scenario has created a barrier for domain experts without computer knowledge and, in particular, expertise in Semantic Web technologies. Indeed, these technologies, despite their potential, are complex and have resulted in a paradoxical situation. On one hand, these technologies have made human intervention more critical due to the semantic interpretation of data that cannot be automated. On the other hand, they have limited the number of curators to those who are experts in the Semantic Web.
This technological advancement has led to two contrasting scenarios in the GLAM sector. Some collections have adopted Semantic Web technologies, requiring more staff with technical expertise for long-term maintenance. Others have refrained from adopting these technologies to avoid curatorial complexities. Examples of these two scenarios have been addressed in the FICLIT Digital Library[1] and OpenCitations [11], two infrastructures handled by the University of Bologna. The FICLIT Digital Library, managed via Omeka S, faces limitations due to its simplistic semantic tools and lack of SPARQL query capabilities, leading to challenges in change tracking and transparent provenance management. In contrast, OpenCitations fully embraces Semantic Web technologies but grapples with the issue of incorrect or missing data, a problem that requires human discernment for correction.
The central problem that arises from these scenarios is the gap between the complex digital technologies and the domain expertise of GLAM professionals. This gap hinders effective data curation and limits the potential of digital collections to represent and disseminate cultural heritage accurately and comprehensively. The solution proposed in this project is the development of a framework that facilitates domain experts without skills in Semantic Web technologies in enriching and editing such semantic data intuitively, irrespective of the underlying ontology model and the technologies adopted for storing such data.
The challenges are manifold. A critical goal is to create a system that is user-friendly for several kinds of end-users, including librarians, museologists, gallery curators, archivists, administrators and IT professionals who are tasked with setting up and maintaining the framework. Another significant challenge is provenance management. In the context of

[1] ADLab - Laboratorio Analogico Digitale and /DH.arc - Digital Humanities Advanced Research Centre, 'FICLIT Digital Library'. University of Bologna, Bologna, Italy, 2022. [Online]. Available: https://dl.ficlit.unibo.it/s/lib/page/home

GLAM institutions, where the historical and source context of data is paramount, a data management system must accurately track and document the responsible agents and primary data sources. Change tracking is also a fundamental requirement. The system needs to efficiently monitor and record all modifications to the data, allowing for transparency and accountability in the curation process. Customization is a further challenge that a data management system for cultural heritage must address. Recognizing that different GLAM domains have unique requirements for how resources are represented and managed, a customizable interface should be tailored to various data models, enabling the representation of diverse resource types according to specific domain needs. Finally, interfacing with pre-existing data presents a substantial challenge, as GLAM institutions often already possess vast collections, which organize their data through the adoption of different data models. This requirement is particularly important for ensuring that the transition to a new data management system is smooth and does not disrupt the ongoing operations of the institution.

2. COMPARATIVE ANALYSIS OF SEMANTIC DATA MANAGEMENT SYSTEMS

The subsequent sections of the paper delve into the specifics of how HERITRACE addresses these challenges. The system's design and functionality are detailed, comparing it with other prominent platforms – i.e. OmekaS [12], Semantic MediaWiki [6], Research Space [9], and CLEF [3] – in terms of user-friendliness, provenance management, change tracking, customization, and data interfacing. These evaluation criteria, employed for the comparative analysis of HERITRACE, are based on those used to assess the CLEF system. This ensures that our assessment criteria are not only relevant but also consistently applied across similar platforms within the digital heritage domain, as summarized in Table 1.

Name	User friendly (Users)	User friendly (Admin)	Provenance Mgmt.	Change-tracking	Customization	Heterogeneous data sources
OmekaS	✓	✓			✓	
Semantic MediaWiki	✓	✓	✓	✓	✓	
Research Space	✓		✓		✓	✓
CLEF	✓	✓	✓	✓		
HERITRACE	✓	✓	✓	✓	✓	✓

Table 1: Comparison of Data Management System Features for the GLAM Sector

OmekaS, recognized for its user-friendly interface, primarily serves museums and educational institutions with its intuitive web-publishing platform. However, it exhibits certain limitations in more complex operational aspects. Notably, OmekaS does not inherently track provenance. This limitation can affect the credibility and traceability of the information presented. Additionally, OmekaS lacks inbuilt change-tracking capabilities. Data interfacing in OmekaS presents another challenge. To import pre-existing data in bulk, users must rely on the CSV Import plugin[2]. This plugin necessitates restructuring the original data to fit its specific format with mandatory field names, which can be a cumbersome and time-consuming process. This requirement for data formatting reduces the platform's flexibility in handling heterogeneous data sources.

Semantic MediaWiki significantly enhances the popular MediaWiki platform by integrating semantic capabilities. This blend of features balances user-friendliness for non-technical end-users and the more complex needs of technical administrators. One of the key strengths of Semantic MediaWiki is its customization potential, although it requires a degree of familiarity with both the MediaWiki environment and underlying semantic concepts. In terms of data provenance management, Semantic MediaWiki provides robust support. However, its capabilities for change tracking are not native to the system but are instead supplemented through the use of external plugins. A notable example is the Semantic Watchlist plugin[3], which effectively monitors changes within the wiki. These changes are stored in a relational database rather than in RDF format, which, while practical for tracking purposes, may not align seamlessly with the semantic structure of the

[2] Berthereau and Corporation for Digital Scholarship, 'CSV Import'. 2015. [Online]. Available: https://omeka.org/s/modules/CSVImport/

[3] WikiTeq, 'Semantic Watchlist'. 2022. [Online]. Available: https://www.mediawiki.org/wiki/Extension:Semantic

data. This discrepancy could potentially restrict the depth of change analysis and the ability to contextualize changes within the semantic framework of the data. Addressing the interfacing with heterogeneous data sources, Semantic MediaWiki initially focused solely on importing OWL ontologies. To broaden its RDF support, the RDFIO extension was introduced [2]. This extension enables the loading of RDF triples, but it is confined to the N-Triples format and notably lacks support for named graphs. This limitation is significant as it restricts the platform's adaptability in various environments that may require more complex semantic data structures.

Research Space, tailored for the academic and research community, excels in user-friendliness for end-users, offering diverse data visualization options such as graphs and temporal maps. However, it maintains a level of complexity for administrators, demanding a steep learning curve. The platform requires a solid understanding of HTML, handlebars and other ResearchSpace-specific components for creating templates, which may be cumbersome even for those with technical expertise. In terms of data provenance, Research Space automatically associates data with its source, ensuring traceability and credibility. However, it lacks a change-tracking system, which could limit its effectiveness in environments where monitoring data modifications over time is crucial. Regarding data interfacing, Research Space allows uploading RDF data directly, which is advantageous for projects involving such formats. However, after the data is uploaded, an administrator's intervention is required to customize the interface appropriately to display the items correctly. This aspect indicates that while Research Space can interface with heterogeneous data sources, doing so involves a significant level of programming complexity for system administrators.

CLEF is designed to manage complex digital libraries, archives, and research data, particularly in the humanities. It offers an administrator-friendly interface and focuses on user-friendliness for end-users, making it suitable for a wide range of audiences within its domain. CLEF's provenance management is robust, utilizing named graphs. Moreover, it does feature change tracking capabilities, including synchronization with GitHub, but lacks a direct system to restore previous versions. Expanding on the capabilities of CLEF, it is important to note that this system does not allow for extensive customization. Moreover, unlike some of its counterparts, CLEF is not designed to upload and manage pre-existing RDF data as-is. This limitation is significant because the software is structured to add items one by one from scratch directly through the user interface. This approach, while potentially beneficial for building new databases, limits the platform’s ability to seamlessly integrate and manage existing large-scale datasets. Furthermore, even though CLEF does not impose a specific data model, it organizes data in a format akin to nanopublications for managing provenance. This structure means that if a pre-existing triplestore is connected, the system is not readily equipped to explore the data without a prior reorganization to make it compatible with CLEF’s framework.

HERITRACE is designed with a focus on usability for domain experts in the fields of archives, libraries, and museums, who may not possess technical skills. Provenance and change management are handled using the OpenCitations Data Model (OCDM) [7], ensuring reliable tracking and documentation of data changes. Lastly, HERITRACE functions seamlessly with RDF data, enabling it to interface with diverse data sources. Additionally, it is user-friendly for administrators responsible for its configuration. The system operates out-of-the-box with RDF data present on any triplestore. For further customization of the user experience, especially regarding data editing forms, HERITRACE utilizes SHACL [10], a well-known language for validating RDF graphs. This approach eliminates the need to learn a specialized language, as with Research Space, while offering a flexibility level that sits between CLEF and Research Space. HERITRACE also allows predefined graphical modifications through YAML [1] configuration files, simplifying the customization process.

In particular, SHACL allows administrators to specify the classes in the data model adopted for describing the data, properties for each class, and constraints for each property. These constraints include the minimum and maximum number of each property for a specific class, the number of permissible values, or the type of values (e.g., class of the value or datatype). Once these SHACL definitions are in place, HERITRACE updates to display editing forms that enforce the constraints defined in the SHACL document.

Furthermore, HERITRACE enables further customization of the interface through a YAML configuration file. This file allows for the definition of user-friendly names for each class and property. It also enables the specification of whether a property should be displayed and, if so, how it should be represented through a SPARQL query. Properties with an inherent order, such as authors, can have their order predicates defined, allowing the interface to present mechanisms for reordering these elements.

HERITRACE automatically manages provenance and change tracking by leveraging the OCDM to ensure meticulous documentation and traceability of data alterations. Each time an entity is created or modified, a new snapshot is generated and stored within a provenance named graph. Classified as `prov:Entity`, these snapshots link to their corresponding entities via the `prov:specializationOf` property. They record essential timestamps, including their creation

(prov:generatedAtTime) and when they become invalid (prov:invalidatedAtTime). The individuals responsible for data changes are documented using the prov:wasAttributedTo property, enhancing accountability and transparency. Crucially, the prov:hasPrimarySource property is employed to trace back to the primary sources of the data, establishing a clear lineage and source of information. This feature is vital for maintaining a continuous historical evolution of each entity. Snapshots are connected to their preceding versions through the prov:wasDerivedFrom property, allowing for a chronological tracking of changes.

Furthermore, the OCDM framework enhances the Provenance Ontology's capabilities by introducing the oco:hasUpdateQuery property. This innovation is pivotal in recording changes to an RDF graph, specifically additions and deletions, through SPARQL INSERT and DELETE queries. This mechanism facilitates the restoration of entities to specific snapshots by reversing operations from all subsequent updates.

HERITRACE's interface also incorporates a timeline feature, as shown in Figure 1, enabling users to explore different versions of the data. This visual representation lets users discern changes between versions at a glance. If a user chooses to restore an entity's previous version, HERITRACE generates a new snapshot. This new snapshot cites the restored snapshot as its primary source, thus maintaining a coherent and traceable record of the data's evolution.

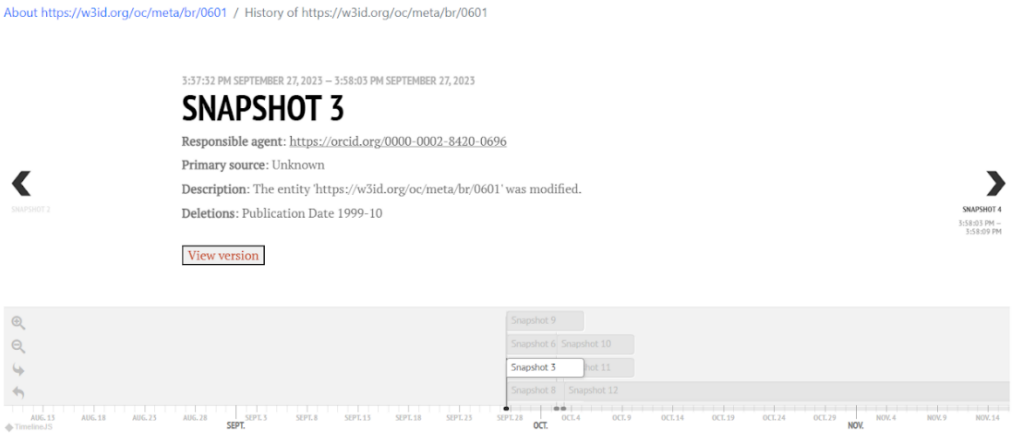


Figure 1. HERITRACE Timeline Interface - This view shows a sequence of data snapshots, allowing users to navigate through versions and view critical metadata for transparent tracking of data provenance and changes.

3. DISCUSSION AND CONCLUSION

Looking towards the future, HERITRACE is poised for significant enhancements. A crucial area of development is the incorporation of an authentication system. This system is vital for ensuring that only authorized personnel can modify metadata, thus maintaining the integrity and credibility of the information. The proposed solution, RCIAM (Identity Access Management for Research Communities) [4], is set to be adopted by the European Open Science Cloud. It will provide a robust framework for user authentication and authorization, leveraging established protocols like OpenID Connect, SAML, and OAuth [8]. This will enable the organization of users into groups, assignment of roles, and management of access rights, enhancing the security and efficiency of data management.

Another pivotal area of development is the integration of RML (RDF Mapping Language) [5] to extend HERITRACE's capabilities beyond native RDF data. This enhancement aims to broaden the system's adaptability to various data formats, particularly tabular formats like CSV and relational databases. The extension of RML is not just about converting different data formats into RDF; it's also about enabling their modification. This advancement is crucial for projects dealing with a wide range of data types, as it will allow for more flexible and comprehensive data handling.

In addition to the future enhancements already outlined for HERITRACE, an important area for further development is the focus on User Experience Insights. Gaining a deeper understanding of how GLAM professionals interact with HERITRACE can provide invaluable feedback for continuous improvement. This involves actively seeking out and analyzing feedback from those who have tested or used the system in real-world scenarios.

Reflecting on the broader implications of HERITRACE's design and functionalities, it is important to consider how such a system aligns with and supports overarching goals of initiatives like the Cultural Heritage Data Space (CHDS) and the European Collaborative Cloud for Cultural Heritage (ECCCH). The CHDS aims to create a unified, accessible, and secure

digital space for European cultural heritage, promoting the sharing and utilization of cultural data across borders. This initiative seeks to enhance the visibility and interoperability of cultural heritage assets, facilitating collaboration among cultural institutions. Similarly, the ECCCH is designed to leverage cloud technologies to foster innovation and collaboration in the cultural heritage sector, providing a platform for sharing resources, tools, and data among cultural institutions and researchers across Europe. Both initiatives underscore the importance of accessibility, interoperability, and collaboration in the digital preservation and dissemination of cultural heritage. While HERITRACE is not directly collaborating with these initiatives, its design principles and functionalities support the shared goals of facilitating access to and collaboration on cultural heritage data, thus contributing to the broader ecosystem of digital cultural heritage management.

In summary, HERITRACE presents itself as a practical solution in the field of semantic data management, with a particular focus on the needs of the GLAM sector. The system provides a user-friendly interface that caters to both non-technical and technical users, alongside features such as provenance management, change tracking, and the ability to customize according to specific needs. Its capability to integrate with existing datasets enhances its practicality. Overall, HERITRACE offers a functional approach to managing digital memory and heritage, potentially contributing to a more comprehensive and accessible understanding of cultural heritage in the digital context. Its design and features position it as a useful tool for professionals in the GLAM sector, aiming to simplify and streamline the management of digital content while respecting the intricacies of cultural heritage data.

For those interested in exploring HERITRACE further, the system along with its documentation are available on GitHub and Software Heritage[4], providing essential resources for implementation and use.

4. ACKNOWLEDGEMENTS

This work has been partially funded by Project PE 0000020 CHANGES - CUP B53C22003780006, NRP Mission 4 Component 2 Investment 1.3, Funded by the European Union - NextGenerationEU.

REFERENCES

[1] Ben-Kiki, Oren, Clark Evans, and Brian Ingerson. “Yaml Ain’t Markup Language (YamlTM) Version 1.1’”. 2005.

[2] Daquino, Marilena, Silvio Peroni, David Shotton, and Arcangelo Massari. “The OpenCitations Data Model.” *Semantic Web Conf. 12507*, 2020. https://doi.org/10.6084/m9.figshare.3443876.v7.

[3] Daquino, Marilena, Mari Wigham, Enrico Daga, Lucia Giagnolini, and Francesca Tomasi. “CLEF. A Linked Open Data Native System for Crowdsourcing.” *ArXiv*, June 1, 2022. https://doi.org/10.48550/arXiv.2206.08259.

[4] De Almeida, Ariovaldo Veiga, Maria Manuel Borges, and Licino Roque. “The European Open Science Cloud: A New Challenge for Europe.” In *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*. New York: Association for Computing Machinery, 2017. https://doi.org/10.1145/3144826.3145382.

[5] Dimou, Anastasia, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. “RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data.” In *Proceedings of the 7th Workshop on Linked Data on the Web*, edited by C. Bizer, T. Heath, S. Auer, and T. Berners-Lee, Vol. 1184. CEUR Workshop Proceedings, 2014. http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

[6] Krötzsch, Markus, Denny Vrandečić, and Max Völkel. “Semantic MediaWiki.” In *The Semantic Web - ISWC 2006*, edited by I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, 4273:935–42. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. https://doi.org/10.1007/11926078_68.

[7] Lampa, Samuel, Egon Willighagen, Pekka Kohonen, Ali King, Denny Vrandečić, Roland Grafström, and Ola Spjuth. ‘RDFIO: Extending Semantic MediaWiki for Interoperable Biomedical Data Management’. J. Biomed Semantics 8, no. 1 (2017): 35. https://doi.org/10.1186/s13326-017-0136-y.

[8] Naik, Nitin, and Paul Jenkins. “Securing Digital Identities in the Cloud by Selecting an Apposite Federated Identity Management from SAML, OAuth and OpenID Connect.” In *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, 163–74. Brighton, United Kingdom: IEEE, 2017. https://doi.org/10.1109/RCIS.2017.7956534.

[9] Oldman, Dominic, and Diana Tanase. “Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace.” In *The Semantic Web – ISWC 2018*, edited by D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, and E. Simperl, 11137:325–40. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-030-00668-6_20.

[4] A. Massari, ‘HERITRACE (Heritage Enhanced Repository Interface for Tracing, Research, Archival Curation, and Engagement)’. Jan. 31, 2024. [Online]. Available: https://archive.softwareheritage.org/swh:1:snp:4d1d83b7043649a21900fcbf6465f0879672228e;origin=https://github.com/opencitations/heritrace

[10] Pareti, Paolo, and George Konstantinidis. “A Review of SHACL: From Data Validation to Schema Reasoning for RDF Graphs.” In *Reasoning Web. Declarative Artificial Intelligence*, edited by M. Šimkus and I. Varzinczak, 13100:115–44. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-95481-9_6.

[11] Peroni, Silvio, and David Shotton. “OpenCitations, an Infrastructure Organization for Open Scholarship.” *Quant. Sci. Stud.* 1, no. 1 (2020): 428–44. https://doi.org/10.1162/qss_a_00023.

[12] Salarelli, Alberto. “Gestire Piccole Collezioni Digitali Con Omeka.” *Bibliothecae.It* 5, no. 2 (2016): 177–200. https://doi.org/10.6092/ISSN.2283-9364/6393.

Libri e biblioteche tra museabilità e musealizzazione digitale: sogno o realtà?

Nicola Barbuti[1], Mauro De Bari[2]
[1] Università degli Studi di Bari Aldo Moro, Italia – nicola.barbuti@uniba.it
[2] Università degli Studi di Bari Aldo Moro, Italia – mauro.debari@uniba.it

ABSTRACT

Questo studio approfondisce il concetto di museabilità e musealizzazione digitale, con particolare attenzione al patrimonio culturale non convenzionalmente considerato museabile sia nella sua dimensione fisica, che nella dimensione digitale. Nello specifico, sono stati presi in considerazione i beni librari e documentali, nell'ottica di valutare modalità di accesso e interazione in ambiente digitale che ne favoriscano la valorizzazione presso un ampio pubblico. La riflessione ha preso avvio dalla valutazione di alcune recenti esperienze di *phygitalization*, che integrano nei beni e nelle collezioni analogiche soluzioni digitali ideate, progettate e realizzate in prospettiva *user-oriented*, espandendo interattivamente l'esperienza di fruizione degli utenti. Questo metodo, infatti, consente di trasformare i beni e le collezioni originali sia fisiche, sia digitali in reti di relazioni articolate e dinamiche, che di fatto ne ampliano il valore informativo e cognitivo in misura incrementale senza limiti di spazio, favorendo nel contempo il coinvolgimento partecipativo degli utenti nell'interazione con le entità di loro interesse. Il metodo è stato sperimentato in due progetti svolti in biblioteche scolastiche con risultati che, allo stato dell'arte, consentono di rivalutare le collezioni librarie e documentali quali beni pienamente museabili, se opportunamente gestiti con soluzioni digitali che ne espandano l'interazione e le ricadute informative e cognitive oltre il contenuto rappresentato.

PAROLE CHIAVE

Museabilità; musealizzazione; digitizzazione; EXEBook; phygitalization.

1. MUSEABILITÀ E MUSEALIZZAZIONE: STATO DELL'ARTE

Nel recente dibattito scientifico sulla valorizzazione del patrimonio culturale fervono studi, ricerche e riflessioni sulla possibilità di avvalersi delle opportunità offerte dalla trasformazione digitale (DT) per rinnovare le strategie di valorizzazione dei beni considerati di interesse del grande pubblico[1].

Tipicamente, i beni analogici concettualmente identificati come *museabili* sono quelli le cui manifestazioni formali estetiche e artistiche esercitano sugli utenti un elevato impatto emozionale. Le istanze di *museabilità* trovano la loro manifestazione concreta nella *musealizzazione*[2], il cui obiettivo primario è ricontestualizzare entità straniate dai loro contesti di ritrovamento in nuove collocazioni, che siano rispondenti a funzioni e bisogni del tutto mutati ed estranei a quelli per i quali sono state create. Di fatto, dunque, nella dimensione museale le entità perdono la loro identità connessa alle istanze e funzioni che le hanno caratterizzate originariamente, per assumere il ruolo del tutto inedito di monadi autoreferenziali, svuotate delle connessioni con i contesti originari di cui, invece, dovrebbero essere testimonianza.

Questo accade perché, nella maggior parte degli istituti culturali, i beni sono selezionati, organizzati ed esposti quasi esclusivamente in base a requisiti formali e con approcci poco *responsive*, finalizzati a proporre l'*oggetto/feticcio* piuttosto che l'*oggetto/mediatore*. Scarsa o nulla attenzione è rivolta alle componenti immateriali che rendono gli oggetti e le collezioni significanti e ne rappresentano il valore culturale più autentico, in quanto ne recuperano e raccontano le relazioni che hanno avuto nelle loro originarie dimensioni spazio-temporali e, quindi, ne definiscono l'identità e la storia [5].

In questo si evidenzia il paradosso fondamentale insito nel concetto di *musealizzazione* come tradizionalmente inteso e applicato: nel valorizzare i beni per le caratteristiche formali e il potenziale emozionale si dimentica che, una volta coinvolti, è poi indispensabile offrire agli utenti la possibilità di approfondire la conoscenza degli oggetti, consentendo loro di esplorarne le relazioni con i contesti spazio-temporali originari. Invece, anche a uno sguardo poco attento risulta evidente che gli istituti culturali focalizzano l'attenzione sulla collocazione attuale dei beni in base ai loro requisiti estetici e artistici, soli elementi ritenuti attrattivi per il grande pubblico.

Nel comune pensare e sentire, infatti, le componenti immateriali sono ricollegate alla dimensione effimera[3] di "storie" a basso coinvolgimento emotivo e cognitivo. Nonostante i recenti studi sulle tecniche di *storytelling* abbiano rivalutato il

[1] https://ilgiornaledellarchitettura.com/2021/01/25/i-musei-e-la-sfida-della-digitalizzazione/
[2] https://www.treccani.it/enciclopedia/musealizzazione-virtuale_(Lessico-del-XXI-Secolo)/
[3] https://losbuffo.com/2021/11/12/ha-senso-musealizzare-la-street-art/

racconto del patrimonio culturale quale fattore fondamentale per la conoscenza [2, 6, 9], nella medesima dimensione effimera del racconto è in realtà tradizionalmente ricondotta la maggior parte del patrimonio culturale, rappresentata da beni valutati ed etichettati *a priori* "a basso interesse storico-culturale" e, quindi, classificati come *non museabili* perché privi di quei requisiti formali in grado di toccare le corde emotive nel grande pubblico. Questo "altro patrimonio" è escluso *di default* dalla valorizzazione su ampia scala e, di fatto, possiamo definirlo *desueto*, sebbene sia composto da beni densi di significato storico e culturale. In esso annoveriamo anche i beni librari, per i quali i concetti di esposizione e fruizione tipicamente si esauriscono e si perdono nell'anonimato di scaffalature inaccessibili al pubblico anche se aperte e armadi chiusi.[4]

A riguardo, sono esemplificative, per citare alcune tra le biblioteche più note, l'Old Library del Trinity College[5], la Casanatense di Roma[6], l'Alessandrina sempre di Roma[7], la Bodleiana di Oxford[8]. Questi istituti rappresentano altrettanti archetipi della più conservativa istanza di *musealizzazione* che ostacola la valorizzazione dei beni: di fatto, ormai da tempo sono esclusivamente luoghi di "transizione", dove i visitatori si aggirano quasi sempre spaesati e confusi in ambienti che, se pur densi di oggetti, contenuti e storia, sono del tutto muti, in quanto non vi è possibilità di accedere e interagire con gli spazi e le opere esposte neanche a livello di semplice contatto[9] . Tenere i libri isolati da qualsiasi modalità di fruizione e interazione è un paradosso in antitesi con l'istanza della cultura accessibile a tutti, che non solo ne annulla le funzioni primigenie di fonti informative e cognitive, ma li condanna a un ruolo di suppellettili mute e invisibili anche nelle loro componenti estetiche e artistiche (vd. Fig.1, Fig.2, Fig.3, Fig.4).

Figura 1. Old Library, Trinity College

Figura 2. Biblioteca Casanatense

Figura 3. Biblioteca Alessandrina

Figura 4. Bodleian Libraries

L'avvento della trasformazione digitale e delle opportunità di elaborare nuovi metodi di valorizzazione del patrimonio grazie alle tecnologie virtuali non ha migliorato questo stato dell'arte. L'adozione massiva degli strumenti digitali per

[4] Desport, Sarah. 2020/2021. «How to make an everyday object possible in a museum? In other words, how to make people want to pay to see objects that surround them on a daily basis?» amusearte. https://amusearte.hypotheses.org/files/2021/05/Sarah-Desport.pdf
[5] https://www.tcd.ie/library/old-library/
[6] https://casanatense.beniculturali.it
[7] https://biblioteche.cultura.gov.it/it/biblioteche-pubbliche-statali/visualizza-le-46-biblioteche/biblioteca/Biblioteca-Universitaria-Alessandrina/
[8] https://www.bodleian.ox.ac.uk/home
[9] https://ilgiornaledellarchitettura.com/2021/01/25/i-musei-e-la-sfida-della-digitalizzazione/

promuovere beni e collezioni non è stata supportata dal cambiamento di mentalità necessario a conoscerli e utilizzarli nel modo migliore in prospettiva *user-oriented*.

Come è ben noto e documentato[10], gli istituti librari sono stati tra i primi ad affidare il proprio rilancio alla digitizzazione ed esposizione online in *digital libraries* delle loro raccolte difficilmente accessibili o inaccessibili, nella convinzione di oltrepassare per questa via le barriere alla fruizione tradizionalmente esistenti e di raggiungere finalmente un ampio pubblico attraverso la rete Internet.

Tuttavia, la gran parte dei progetti di digitizzazione di raccolte storiche realizzati negli ultimi quindici anni, anche i più recenti attuati secondo le migliori pratiche e tecniche in uso[11], hanno di fatto generato collezioni di *digital twins* che risultano disinteressanti per il grande pubblico quanto i corrispettivi analogici, se non di più[12]. A nostro parere, l'equivoco è generato dalla presunzione che digitizzare le raccolte librarie sia di per sé sufficiente a dare corpo alle recenti istanze di *musealizzazione virtuale* [4], secondo cui la valorizzazione delle collezioni risiede nella creazione e adozione di soluzioni digitali tridimensionali fruibili anche da remoto. Un equivoco che si riflette nella fruizione quantitativamente e qualitativamente sottodimensionata e quasi sempre occasionale o cursoria sia delle *digital libraries* esposte in rete, sia delle obsolete soluzioni virtuali realizzate dai musei, mostrando chiaramente la sperequazione tra l'ambiziosa portata dei progetti e l'impatto del tutto insignificante rispetto alle attese[13].

Di fatto, questa presunzione ha origine anche nella sottovalutazione dei nuovi bisogni di conoscenza che la DT ha portato con sé, che ha condizionato e compromesso la capacità di progettare e realizzare la digitizzazione secondo prospettive orientate a favorire le nuove interazioni richieste dagli utenti digitali [3]. Le ricadute sono sotto gli occhi di tutti: anche le generazioni native digitali, che dovrebbero essere mentalmente predisposte a interagire con qualsiasi entità fruibile tramite dispositivo digitale, rientrano nel quadro di disinteresse generale per i progetti di digitizzazione, fatta eccezione per quelle nicchie di utenti [7] afferenti soprattutto ai contesti della *gamification*.

2. MUSEABILITÀ E MUSEALIZZAZIONE: UN RIPENSAMENTO NECESSARIO

Partendo dallo stato dell'arte sopra delineato, in maniera funzionale riteniamo sia oggi irrinunciabile e indispensabile innanzitutto ripensare i concetti di *museabilità* e *musealizzazione* riconducendoli alla dimensione digitale, unica in cui è possibile rigenerare e rendere disponibile sia il patrimonio stereotipizzato, sia l'"altro patrimonio" valorizzando non solo i requisiti estetici e artistici degli oggetti, ma anche e soprattutto quelli immateriali, culturali e storici che ne ricostruiscono le identità originarie.

A livello teorico, infatti, è nostra convinzione che i requisiti di *museabilità digitale* di un'entità culturale risiedano innanzitutto nel *racconto* che essa può esprimere e manifestare: quando identificato e materializzabile in ambiente digitale, esso la valorizza non più solo di per sé, ma soprattutto nelle relazioni con i contesti originari cui può essere ricondotta e relazionata.

Conseguentemente, si impone il ripensamento e la rivalutazione dei beni da considerare museabili: inevitabilmente, non possono più essere solo quelli che riscontrano standard o criteri formali ed estetici, ma, in coerenza con gli obiettivi più ampi di identificazione, tutela, promozione e valorizzazione del patrimonio culturale, sono museabili tutte le entità portatrici di *significanti* che, in quanto tali, sono mediatori di nuova conoscenza ed esperienza verso chi ne fruisce.

Quindi, anche l'intero "altro patrimonio" *desueto* ha piena dignità di essere valorizzato in quanto digitalmente museabile, e le raccolte librarie storiche sia già esposte, sia di prossima esposizione in digital libraries possono essere oggetto di progetti di *musealizzazione virtuale*. Anche questa va affrancata dallo stereotipo identificativo "soluzione 3D" e ridefinita quale processo che miri a riconnettere in ambiente digitale qualsiasi entità o collezione alle identità originarie, rigenerando contesti nei quali è possibile materializzarne le relazioni, i messaggi e il modo con cui si intende manifestarli.

Dunque, nel valutare la museabilità e procedere con una musealizzazione virtuale che applichi questo approccio diventa fondamentale tenere conto della complessità e molteplicità di fattori che caratterizzano e definiscono culturalmente ciascuna entità, quali il suo originario contesto di provenienza, la sua storia precedente, le testimonianze sul suo ciclo di vita, la necessità di ridefinirne le funzioni nelle relazioni con i nuovi utenti e con i contesti di collocazione, al fine di garantirne la percezione corretta nelle sue rinnovate funzioni all'interno della dimensione museale digitale.

[10] Valga qui limitarsi a riportare per le biblioteche quanto fruibile in Internet Culturale (https://www.internetculturale.it/) e in Alphabetica (https://alphabetica.it/web/alphabetica/); per gli archivi, citiamo il progetto del SAN (https://san.beniculturali.it/web/san/progetti-di-digitalizzazione), sebbene non sia più aggiornato da tempo.

[11] Tra le DL attualmente esposte in rete che presentano elementi di maggiore interesse ai fini dell'applicazione di istanze di musealizzazione si segnalano: https://dl.bnnonline.it/handle/20.500.12113/4758; ETC. NAPOLI (sempre 4Science); BAV

[12] https://digitalcommons.cwu.edu/cgi/viewcontent.cgi?article=1137&context=libraryfac

[13] https://www.treccani.it/magazine/lingua_italiana/articoli/parole/Figitale.html

Se questo metodo fosse applicato a libri e biblioteche, siamo certi che ne incrementerebbe il potenziale di conoscenza e di valorizzazione anche presso un ampio pubblico. Del resto, una biblioteca che non stimola gli utenti a interagire con le raccolte, rendendole inavvicinabili anche solo per leggere i dorsi dei libri esposti sugli scaffali, equivale a un'entità senza più alcuna storia da raccontare oltre quella claustrofobica dello spazio in cui ci si muove senza meta, e quindi è destinata prima all'entropia, poi all'obsolescenza[14].

Tuttavia, anche le più innovative esperienze di *musealizzazione virtuale* che hanno recepito e applicato queste istanze non hanno migliorato l'impatto sugli utenti: persistono, infatti, fruizione scarsa ed evidente disinteresse riconducibili all'intangibilità delle soluzioni, che rappresenta una criticità imprevista e una barriera all'accesso ancora oggi irrisolte. È evidente che anche la mancanza di "vicinanza fisica" è un fattore disincentivante, soprattutto se il gemello digitale si presenta basicamente statico e piatto quanto la matrice analogica.

Alcuni recenti progetti di ricerca sperimentale hanno forse segnato una svolta a questo impasse e inaugurato una nuova stagione della DT, in cui fisico e digitale si integrano in processi di *phygitalization*[15] [4]. Il metodo *phygital* si propone di trasformare gli oggetti e i luoghi fisici in accessi a contenuti digitali che ne arricchiscono il racconto e sono fruibili tramite dispositivi mobili, nella prospettiva di stimolare gli utenti ad approfondirne la conoscenza [8]. Essi, quindi, possono interagire attivamente e dinamicamente con l'entità fisica senza che sia necessario altro contatto diretto se non la prossimità fisica, mantenendo la libertà di muoversi nell'ambiente. Tra i diversi progetti realizzati, valga qui ricordare l'esperienza *Immersive Dickens*[16] del Victoria and Albert Museum (V&A)[17], che ha trasformato in entità *phygital* alcuni manoscritti di Charles Dickens conservati presso il museo[18] , integrandoli con soluzioni digitali che hanno attivato nuove possibilità di interazione da parte degli utenti[19] , riuscendo a coinvolgere partecipativamente anche un pubblico giovane. La soluzione rappresenta una felice applicazione di nuove istanze di *musealizzazione phygital* applicate a un bene *desueto*, di fatto certificandone la museabilità e l'orientamento verso prospettive *user-oriented*.

3. MUSEALIZZAZIONE PHYGITAL E BIBLIOTECHE, UNA PROSPETTIVA REALE

Partendo dallo studio di alcune esperienze pilota [1], si è ritenuto utile sperimentare il potenziale della *musealizzazione phygital* all'interno di due biblioteche scolastiche, nell'ottica di valutarne l'impatto sull'interesse di giovani studenti di scuola secondaria superiore di età compresa tra i 16 e i 18 anni[20]. La prima sperimentazione è stata realizzata nel 2019 nella biblioteca del Liceo "Francesco de Sanctis" di Trani[21], nell'ambito del progetto PON *Comunicare e promuovere un evento culturale*. Il secondo è stato realizzato nel 2022 presso la biblioteca del Liceo "Ignazio Vian" di Bracciano[22] per il progetto PCTO *Percorsi per le competenze trasversali e per l'orientamento*.

Gli studenti sono stati raggruppati in gruppi di lavoro di massimo cinque persone. Ciascun gruppo è stato preliminarmente istruito sulle fasi del processo di *phygitalizazion*, che poi hanno applicato su alcuni volumi da loro scelti tra quelli disponibili in biblioteca. La selezione ha privilegiato testi comunemente ritenuti di scarso o nullo interesse: raccolte di documenti diplomatici medievali, trattati illustrati di moda in lingua inglese della fine dell'Ottocento, saggistica storica e storico-artistica, poesia antica o contemporanea, manuali di fisica applicata, e così via. Gli studenti hanno si sono poi impegnati nella figitalizzazione dei volumi, eseguendo le attività necessarie ad arricchirli di contenuti digitali da rendere fruibili tramite dispositivi mobili:

- predisposizione dei marker come pulsanti per l'accesso a hyperlink tramite TUI, utilizzando le funzionalità della piattaforma *Genial.ly*[23];
- selezione dei testi e markup delle parti di interesse individuate;
- scelta e selezione, o creazione dei contenuti digitali con cui arricchire il testo;
- linking dei contenuti digitali ai marker;
- predisposizione delle funzionalità *responsive* dei volumi phygital per la fruizione su dispositivo mobile;

[14] https://www.istat.it/it/files//2023/05/STATISTICA_TODAY_Libri_biblioteche.pdf
[15] https://www.treccani.it/magazine/lingua_italiana/articoli/parole/Figitale.html
[16] https://www.vam.ac.uk/research/projects/immersive-dickens
[17] https://www.vam.ac.uk
[18] https://gtr.ukri.org/projects?ref=AH%2FR010064%2F1
[19] https://www.vam.ac.uk/blog/digital/how-can-technology-improve-the-museum-experience
[20] Il primo progetto è stato realizzato nel 2019 con il Liceo "Francesco de Sanctis" di Trani nell'ambito del progetto PON *Comunicare e promuovere un evento culturale*, e ha coinvolto 20 studenti dei trienni e alcuni loro docenti; il secondo è stato realizzato nel 2022 con il Liceo "Ignazio Vian" di Bracciano per il progetto PCTO *Percorsi per le competenze trasversali e per l'orientamento*, e ha coinvolto 29 studenti dei trienni.
[21] https://w3.liceodesanctis.edu.it/struttura/liceo-statale-classico-linguistico-scienze-umane-francesco-de-sanctis/
[22] https://www.liceovian.edu.it
[23] https://genial.ly/it/

- caricamento in ambiente di test;
- test di accessibilità alle espansioni e delle funzionalità di interazione utente: gli studenti hanno verificato che, inquadrando i testi marcati con la fotocamera dei propri dispositivi mobili e scansionandoli, il display mostrasse la riproduzione della pagina con evidenziati i marker di accesso alle estensioni e le informazioni sui tipi di contenuti accessibili, utilizzabili tramite TUI.

Per l'integrazione delle soluzioni digitali nei testi e lo sviluppo delle funzionalità di interattività è stata utilizzata l'app *EXpanded Endless Book* (EXEBook)[24].

4. CONCLUSIONI E PROSPETTIVE

Dopo quanto detto, riteniamo che l'approccio *phygital* rappresenti nell'ecosistema culturale il corto circuito in grado di riattivare il nesso virtuoso beni-utenti-interazione-conoscenza. I risultati delle sperimentazioni, infatti, hanno mostrato che l'adozione della *musealizzazione phygital* in biblioteca sortisce impatti e ricadute positivi su un duplice livello. In primo luogo, libri *desueti* e comunemente considerati privi di qualsiasi appeal sia per i contenuti, che per la forma sono diventati oggetti di interesse per il loro potenziale, prima inesplorabile, di arricchirsi di contenuti digitali e trasformarsi in entità culturali dinamiche e autenticamente interattive, in grado di attivare l'interesse anche di utenti giovani.

In secondo luogo, materializzando in ambiente digitale storie, narrazioni e contenuti altrimenti non fruibili in alcun modo, i volumi *figitalizzati* nelle due biblioteche scolastiche si propongono quali promettenti archetipi del valore della *musealizzazione phygital* quale processo che, finalmente, materializza in pieno le istanze di museabilità in grado di restituire ai beni la loro vocazione culturale più autentica.

Resta aperta la sfida di trasformare questo approccio, ancora sperimentale e occasionale, in una buona pratica da disseminare all'interno di tutte le istituzioni non solo bibliotecarie, nella prospettiva di intraprendere strategie di valorizzazione che rendano oggetti e collezioni *attivatori di interesse* per un pubblico sempre più ampio, restituendo nel contempo anche ai beni desueti la loro dignità culturale e storica.

Ma questa prospettiva rende necessaria un'altra e altrettanto densa riflessione sul ruolo cruciale che i professionisti del patrimonio culturale hanno, nella sfida di confrontarsi con un'innovazione che richiede padronanza sia del metodo, che delle tecniche necessarie ad applicarlo nel modo migliore. Si rende necessario un vero e proprio *mindset change*, un cambiamento di mentalità in direzione della rivalutazione della digitizzazione quale vero e proprio metodo di generazione di ecosistemi culturali che, di fatto, sono la manifestazione della trasformazione che stiamo vivendo. Questo implica la necessità per i professionisti culturali di muoversi in perfetto equilibrio tra tradizione e cambiamento, adottando un approccio all'innovazione aperto e flessibile che permetta loro di affrontare consapevolmente le sfide e le opportunità della *phygitalization*.

BIBLIOGRAFIA

[1] Andrade, José Gabriel, e Patricia Dias. «A phygital approach to cultural heritage: augmented reality at Regaleira». *Virtual Archaeology Review* 11, fasc. 22 (2020): 15–25.

[2] Bonacini, Elisa. «Storytelling digitale in ambito culturale e il suo ruolo in ambito educativo». *Culture Digitali* 1 (2021): 85–101.

[3] Bonn, Anna, Francis Saa-Dittoh, e Hans Akkermans. «Bridging the digital divide». In *Introduction to Digital Humanism. A textbook*, a cura di Hannes Werhner, Carlo Ghezzi, Jeff Kramer, Julian Nida-Rümelin, Bashar Nuseibeh, Erich Prem, e Allison Stanger, 283–98. Springer Cham, 2024. https://doi.org/10.1007/978-3-031-45304-5_19.

[4] Gamper, Christian. «Ambienti digitali e sviluppo dell'audience nei musei. Monitoraggio dell'esistente e analisi delle potenzialità». *Culture Digitali* 0 (2021).

[5] Giannini, Tula, e Jonathan P. Bowen, (a cura di). *Museums and Digital Culture. New Perspectives and Research*. Springer, 2019.

[6] Hutson, James, e Piper Hutson. «Storytelling». In *Inclusive Smart Museums: Engaging Neurodiverse Audiences and Enhancing Cultural Heritage*, a cura di James Hutson e Piper Hutson, 49–84. Palgrave Macmillan Cham, Springer Nature Switzerland, 2024.

[7] Lu, Shan Shan, Ruwen Tian, e Dickson K.W. Chiu. «Why do people not attend public library programs in the current digital age? A mix method study in Hong Kong». *Library Hi Tech* 42, fasc. 4 (2024): 1237–65. https://doi.org/10.1108/LHT-04-2022-0217.

[8] Mele, Cristina, Tiziana Spena Russo, Marialuiza Marzullo, e Irene Di Bernardo. «The phygital transformation: a systematic review and a research agenda». *Italian Journal of Marketing*, 2023, 323–49. https://doi.org/10.1007/s43039-023-00070-7.

[9] Mele, Francesco. «Ricostruzioni virtuali e digital storytelling per la valorizzazione di contesti artistici perduti: Case studies e modelli a confronto». *DigItalia* 18, fasc. 1 (2023): 198–203. https://digitalia.cultura.gov.it/article/view/2998.

[24] https://www.youtube.com/watch?v=3hibrvCSI5M

Listening2Painting: an Audio Augmented Reality approach for Arts

Nicola Orio[1], Daniel Zilio[2], Andrea Micheletti[3]
[1] Department of Cultural Heritage, University of Padua, Italy - nicola.orio@unipd.it
[2] Department of Cultural Heritage, University of Padua, Italy - daniel.zilio@unipd.it
[3] Department of Cultural Heritage, University of Padua, Italy - andrea.micheletti@unipd.it

ABSTRACT
This paper describes the first stage of the Listening2Painting project, which involves researching the effects of experiencing paintings through sonification. The report includes details of a pilot study and the initial version of an application as part of a larger project that involves multiple university students.

KEYWORDS
Audio-Augmented Reality (AAR); Arts; Mobile development.

1. INTRODUCTION
Painting has been one of the most effective and treated forms of communication and expression since the origins of the human species. Visiting art museums and observing the marvelous artworks of the past and present is a hallmark of cultural tourism. However, due to a lack of knowledge or limited time, we often fail to fully comprehend the elements artists intended to communicate through their works. A relevant contribution could be offered by using technologies such as Audio-Augmented Reality (AAR) to reduce this distance between the observers and the paintings. Utilizing the definition provided in [6] the AAR presents an additional layer of contextual information in addition to the user's experience with the real world, and the additional layer is presented in audio form. In this paper, we present the initial stage of the project "Listening2Painting". The main objective of this research is to explore how a tool like AAR can be utilized in the art world to ensure that a wider audience can comprehend the meaning of a painting. The study encompasses several areas of interest, including enhancing art appreciation and understanding, improving the museum experience for visitors, and introducing an innovative approach to teaching computer science skills to humanities students.
A milestone point is the release of a mobile application called L2P, allowing users to interact with a collection of selected artworks in various ways. This app will provide a unique experience by enabling users to explore several paintings using, at the same time, hearing and sight senses. The project aims to involve users in each step of the designing phase, from creating a participatory audio-labeled collection of artworks to being testers and peer-reviewers of the community work.
The project's objective can be summarized in three points. Firstly, to enhance the visitor's experience at the museum. Secondly, to provide a tool that enables the reading of a painting by uncovering details and making it easier for visitors to understand the choices made by the painter. Finally, the project aims to serve as a way to teach computer skills to humanities students. There are several ways to enrich the experience of exploring an artwork through audio. In [4] it is described as an interactive system that uses machine learning to recognize objects inside Claude Monet’s painting automatically. In this research, the authors manually created a training dataset and enriched the user experience with soft music and natural sounds played in response to mouse positioning. The automatic sonification of a group of four artworks is the topic discussed in [3]. The music is produced using a designated algorithm [5] and further developed by a musician. Finally, a laboratory test is carried out in which participants view the reproduction of the painting while listening to the proposed audio in a setting that mimics a museum gallery. A sentiment analysis was conducted to explore participants' reactions. The approach proved useful in enhancing their experience. The effects of background music on the aesthetic experience of visual art are presented in [8]. In this paper, the authors explore the emotional impact of background music while observing an abstract painting by Wassily Kandinsky, reporting the effects on the experience reported by a group of visitors and how affected their judgments about the artwork. The use of sound in pictorial observation is also a topic of interest in the case of people who are blind or visually impaired, and there are several different approaches to this task. For instance, in [2] the relationship between sound and color is analyzed. The goal is to codify a set of colors with different melodies that enhance the experience for people with visual impairments. The touchscreen exploration and the verbal feedback are investigated in [1], where two different approaches are presented for segmenting elements within a painting for presentation. Finally, proxemic audio is the approach involved in [7]. Using a Microsoft Kinect to detect the observer's position and so provide

specific audio. In this paper, we first present a pilot study, followed by the web platform where the first group of participants contribute, and finally, the alpha version of the app.

2. PILOT STUDY

A prototype version of the L2P has been created to enable visitors of the "Palazzo Chiericati" Museum in Vicenza to interact with four selected paintings. Each painting has been associated with the most significant elements which are represented through sound. For instance, the unsheathed sword is represented through the bright noise, the calm babble of a river represents the house of Nereids at the court of Diana, and the sound of wind strokes on a flag represents the banner of an army. An attempt was made to reproduce the sounds in such a way that they were as representative as possible of the symbolic representation of the work, thus taking into account the impact of the individual elements within the work itself. The choice of sounds and their processing was made not by a stereophonic expert but by an art history expert as part of a thesis project. The relevance of the pictorial element has guided the choice of these elements. The aim was to make visitors more participative than simply contemplative. However, it's not always possible to associate all the relevant elements with a sound, and not all of the elements that produce sounds are relevant to the picture. After completing the acoustic description, the sounds have been added to the interactive app. A tablet was positioned in front of each painting and visitors were invited to listen either to a soundscape that presented all the sounds in a sequence that suggested the direction in which the painting was supposed to be read or to interact with the touch screen that presented an image of the painting, activating the corresponding sound when the element was touched. Twelve museum visitors, aged between 18 and 30, freely agreed to participate in the test. They were provided with headphones to ensure that they could perceive the sounds well and not disturb other museum visitors. The participants were divided into three different groups representing three different routes, this one in a random manner, alternating between listening to the sound narration, interacting with the app, or without any sounds. Results have been encouraging. In a short interview carried out after the experience, visitors showed appreciation for the experience, saying that it was involving and informative at the same time. No further information was provided in this initial prototype, so the user experience was limited to listening to the sounds. However, some of the users expressed an interest in knowing more about the reason why some elements have been sonified. An interesting outcome of the interviews was that visitors slightly preferred the passive experience of just listening to the soundscape because they were afraid of missing some sonic elements through the interaction.

3. THE WEB PLATFORM AND THE APP

The next step in our research is to transform the prototype into a tool that allows anyone to create an AAR experience using the app by selecting the painting and adding the sonic elements.

At this initial stage, a group of about twenty students from the bachelor's degree program in Cultural Tourism Planning and Management and Art History were involved in the project. The first operation is the sonification of a proposed set of artworks. Students are asked to select one painting, without limitations about artistic movement or genre, and manually set the opera pieces in which users can interact. After the mandatory registration, the web portal provides a page dedicated to the insertion of the artwork. Figure 1 shows on the left the parameters required: Title, Artist, Year, a brief presentation of the painting, and two multimedia files, respectively the image file (in PNG or JPG/JPEG) format and file audio, which is optional in this stage of the project, which represent the complete soundscape. The resulting page is on the right side of Figure 1. Users can modify each field whenever they want.

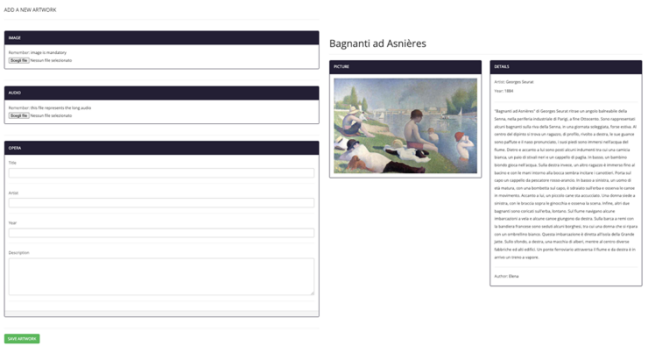


Figure 1. Screenshot of the web Portal

Once the fundamental element of the artwork is inserted, users can proceed with the process of sonification. To accomplish this task, a dedicated webpage has been created. The required information for the sonification process includes the audio file, the rectangle coordinates associated with the sound, the volume level, and a brief description. A flag indicates whether the sounds should be played in the background without user interaction to provide a basis on which individual sounds are added: the sound of wind through the leaves in a forest, the voices of a crowd of people, and so on.
Participants added to the Google Android Dashboard tester group can now test and check the added elements. The app is currently being tested internally and is only available on Android. The aesthetics have been kept to a minimum for testing purposes. Figure 2 and Figure 3 show, respectively, the dynamic list with all artworks and the basic screenshot of one of them. Users are guided to identify rectangles in red shadows. We plan to remove this function in the final version. Users can explore the painting with their fingers and play linked sounds automatically.

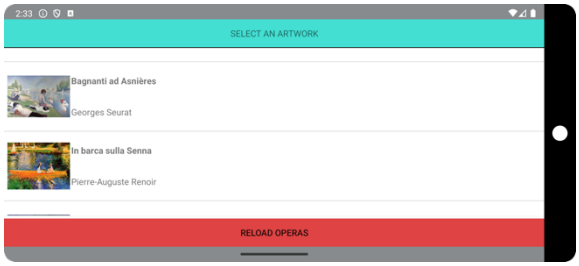


Figure 2. App screenshot

Figure 3. App screenshot

4. FUTURE WORKS

As reported, this work is in the initial stage. There are several directions in which the research can be developed. First of all, we plan to improve the interaction by providing information about the elements associated with sounds. Once the visitor's attention is directed towards an element, we can provide a description of that element, its role in the picture, the artistic choices of the painter, and so forth. Moreover, additional feedback can be provided about the number of sonic elements, with a counter highlighting how many elements have been selected. Most of all, we intend to replicate the initial experiment in a real museum with a larger group of users testing the L2P app and providing feedback.

REFERENCES

[1] Ahmetovic, Dragan, Nahyun Kwon, Uran Oh, Cristian Bernareggi, and Sergio Mascetti. 'Touch Screen Exploration of Visual Artwork for Blind People'. In *WWW '21: Proceedings of the Web Conference 2021, April 19–23, 2021, Ljubljana, Slovenia*, edited by IW3C2 (International World Wide Web Conference Committee), 2781–91. ACM, 2021. https://doi.org/10.1145/3442381.3449871.

[2] Cho, Jun Dong, Jaeho Jeong, Ji Hye Kim, and Hoonsuk Lee. 'Sound Coding Color to Improve Artwork Appreciation by People with Visual Impairments'. *Electronics* 9, no. 11 (2020). https://doi.org/10.3390/electronics9111981.

[3] Dam, Abhraneil, Yeaji Lee, Arsh Siddiqui, Wallace Santos Lages, and Myounghoon Jeon. 'Enhancing Art Gallery Visitors' Experiences through Audio Augmented Reality Technology'. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2023*, edited by Dick de Waard, Vera Hagemann, Linda Onnasch, Antonella Toffetti, Denis Coelho, Assaf Botzer, Marco De Angelis, Karel Brookhuis, and Stephen Fairclough, 67 (1):971–77, 2023. https://doi.org/10.1177/21695067231192706.

[4] Gayhardt, Lauryn, and Margareta Ackerman. 'SOVIA: Sonification of Visual Interactive Art'. In *Proceedings of the 12th International Conference on Computational Creativity (ICCC '21). Mexico*, edited by Broad Terence, Berns Sebastian, Colton Simon, and Grierson Mick, 2021.

[5] Manaris, Bill, Blake Stevens, and Andrew R. Brown. 'JythonMusic: An Environment for Teaching Algorithmic Music Composition, Dynamic Coding and Musical Performativity'. *Journal of Music, Technology & Education* 9, no. 1 (2016): 33–56.

[6] Naphtali, Dafna, and Richard Rodkin. 'Audio Augmented Reality for Interactive Soundwalks, Sound Art and Music Delivery'. In *Foundations in Sound Design for Interactive Media*, edited by Michael Filimowicz, 300–332. New York: Routledge, 2019. https://doi.org/10.4324/9781315106342-14.

[7] Rector, Kyle, Keith Salmon, Dan Thornton, Neel Joshi, and Meredith Morris Ringel. 'Eyes-Free Art: Exploring Proxemic Audio Interfaces For Blind and Low Vision Art Engagement'. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2017.

[8] Thenille, Braun J., Bruna de Oliveira, Giulia Ventorim Ferreira, João R Sato, Cláudia Feitosa-Santana, and Patricia Vanzella. 'The Effect of Background Music on the Aesthetic Experience of a Visual Artwork in a Naturalistic Environment'. *Psychology of Music* 51, no. I (2023): 16–32.

Luoghi comuni: metodi e strategie di sviluppo software in ambito GLAM, dalle voci di autorità all'esplorazione cartografica

Herbert Natta[1], Gianluca Rossi[2], Roberta Maggi[3]
[1] Istituto di Matematica applicate e tecnologie informatiche IMATI-CNR, Italia - herbert.natta@ge.imati.cnr.it
[2] Istituto di Matematica applicate e tecnologie informatiche IMATI-CNR, Italia - gianluca.rossi@ge.imati.cnr.it
[3] Istituto di Matematica applicate e tecnologie informatiche IMATI-CNR, Italia - roberta.maggi@cnr.it

ABSTRACT
L'interoperabilità dei sistemi informativi, in ambito umanistico e non solo, rappresenta un obiettivo e una sfida che coinvolge, a diversi livelli, progetti di sviluppo software in numerosi settori di applicazione. In particolare, gli strumenti digitali impiegati nella gestione, archiviazione, descrizione e pubblicazione di cataloghi di beni culturali (archivi, musei, biblioteche) si stanno evolvendo nella direzione, da un lato, di integrare standard, formati, tecnologie che agevolano la condivisione e interconnessione di dati, richiedendo quindi, dall'altro lato, l'attivazione di protocolli di comunicazione tra domini e ambiti disciplinari distinti. Questa trasformazione è centrale per i software che operano in ambito GLAM, per i quali la gestione della trasversalità cross-dominio richiede soluzioni, tecniche e scientifiche, che interessano i diversi livelli dell'applicativo (dalla struttura della base dati alle modalità di presentazione). Inscrivendosi in questo filone di ricerca, il contributo presenta le scelte, tecnologiche e metodologiche, operate nell'evoluzione di Geca, software per la gestione di cataloghi di beni culturali, con particolare attenzione i) all'identificazione delle entità e connessioni semantiche che, anche in una base dati strutturata su un modello logico relazionale, costruiscono una rete di punti di accesso utile ii) alla progettazione e sviluppo di strumenti e modalità di presentazione ed esplorazione integrata dei dati in catalogo, fino a ipotizzare iii) il livello di presentazione come servizio autonomo, modulare e personalizzabile. Si presenta qui la metodologia adottata, le soluzioni tecniche e architetturali, insieme ai primi risultati ottenuti e agli sviluppi previsti.

KEYWORDS
GLAM; interoperabilità; open data; authority file; cartografia.

1. INTRODUZIONE
La gestione informatizzata di cataloghi di beni culturali, e la condivisione dei relativi dati e metadati, richiede la progettazione di strumenti software capaci, da un lato, di garantire la granularità e la specificità disciplinare dei tracciati descrittivi e, dall'altro, di strutturare e presentare i dati in modo da agevolare l'accesso al loro contenuto informativo e l'interscambio con altri sistemi informatici. L'interoperabilità tra strumenti digitali è infatti un nodo cruciale nel processo di digitalizzazione della conoscenza, necessario a evitare effetti collaterali prodotti dall'accumulo di dati non sistematizzati. In questo senso, la definizione e diffusione di linee guida per la pubblicazione di dati relativi ai beni culturali ispirate ai principi FAIR (Findable - Accessible - Interoperable - Reusable, cioè reperibili, accessibili, interoperabili e riutilizzabili) [13] e al concetto di dato come bene comune definisce un framework teorico, metodologico e normativo utile a orientare la progettazione e sviluppo di strumenti digitali per la raccolta, archiviazione e condivisione di dati, in ambito culturale e non solo[1].
Se l'adozione di tecnologie e modelli utili a strutturare l'informazione per la sua condivisione è fondamentale per il dialogo *inter*-sistema, la centralità di questa strategia di sviluppo emerge con maggiore evidenza nella costruzione di relazioni *intra*-sistema, requisito dei software che ambiscono a gestire cataloghi di beni culturali eterogenei [3]. La sfida aperta in ambito GLAM è infatti quella di mettere in rete conoscenze accumulate in domini e settori disciplinari diversi, supportando sia le esigenze operative di enti e istituzioni attive trasversalmente, in forma centralizzata o distribuita, nella gestione del patrimonio culturale, sia quelle informative attivate da linee di ricerca sempre più orientate alla *trans*- e *inter*-disciplinarità. A questo si aggiunge, come aspetto non secondario, l'orizzonte aperto dalla diffusione e applicazione in diversi ambiti

[1] Si vedano a questo proposito la definizione di *open* proposta dalla Open Knoledge Foundation (https://opendefinition.org/od/2.1/en/) e, nel contesto nazionale, la recente pubblicazione delle "Linee Guida recanti regole tecniche per l'apertura dei dati e il riutilizzo dell'informazione del settore pubblico" (https://www.agid.gov.it/sites/default/files/repository_files/lg-open-data_v.1.0_1.pdf), risultato di un processo collaborativo che ha visto impegnata l'Agenzia per l'Italia Digitale, insieme a enti di ricerca e istituzioni pubbliche, ma anche, a livello europeo, il "Data Act" (http://data.europa.eu/eli/reg/2023/2854/oj), esito operativo di una più generale strategia europea per i dati (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066).

disciplinari di modelli di intelligenza artificiale, il cui sviluppo trae vantaggio dalla disponibilità di dati e dalla loro interconnessione [12].
In questo contesto, si propone quindi come caso studio, utile a condividere scelte e criticità, lo sviluppo del software Geca [6], originariamente progettato per la gestione di un catalogo collettivo di risorse bibliografiche e recentemente esteso a includere la descrizione di beni archivistici e, con riferimento agli standard catalografici ICCD, oggetti d'arte, fotografie e beni relativi al patrimonio scientifico-tecnologico[2].
In particolare, il contributo intende presentare il processo di revisione della struttura dati delle voci di autorità come innesco di una trasformazione del sistema informatico volta a identificare i punti di contatto tra domini e a enfatizzarne le relazioni semantiche, aprendo a modalità integrate di presentazione ed esplorazione dei dati.
All'interno di questa linea evolutiva, si evidenzia nello specifico il ruolo dei luoghi come entità utili, da un lato, ad aprire punti di accesso trasversali alle diverse tipologie di risorse in catalogo e, dall'altro, come possibilità di far emergere relazioni spaziali tra i dati e, conseguentemente, nuove modalità di esplorazione [11].
L'informazione spaziale è un elemento fondamentale nell'informatizzazione della gestione dei beni culturali [7, 2] in diversi ambiti, dalle applicazioni più strettamente amministrative (gestione del posseduto, prestiti interbibliotecari, ecc.) [5] a quelle che integrano nell'attività di catalogazione funzionalità di geolocalizzazione e notazione spaziale delle risorse fino allo sviluppo di web-gis e web-map destinate a fornire nuovi punti di accesso e visualizzazione dati [10, 8].
L'acquisizione e rappresentazione del dato spaziale ha infatti una doppia funzione nella progettazione di un sistema informativo per i beni culturali: da un lato un'orizzontalità che trasversalmente interconnette domini diversi attraverso un comune sistema di riferimento, quello cartografico, e, dall'altro, una verticalità che attraversa i diversi livelli dell'architettura predisponendola a modalità di fruizione e interrogazione che restituiscono all'esperienza dell'utente i dati in una forma arricchita di un contenuto informativo latente, determinato appunto dalle loro relazioni spaziali.
In quest'ottica di trasversalità bidirezionale, la trasformazione del software ha pertanto interessato non solo la base dati, ma anche il livello di presentazione, con lo sviluppo di una soluzione di frontend personalizzabile, progettata secondo il paradigma FEaaS (Front End As A Service) [4].
L'estensione degli ambiti di applicazione del software richiede una flessibilità coerente con la varietà di utenti con i quali è chiamato a interagire, incoraggiando un approccio modulare che, grazie all'introduzione di un livello middleware che garantisca stabilità nell'esposizione dei dati, permetta di massimizzarne l'adattamento a diversi scenari di utilizzo.
Si descrivono di seguito, più nel dettaglio, le principali caratteristiche delle tre linee evolutive sviluppate (revisione del modello dati delle voci di autorità, introduzione del modulo di esplorazione cartografica e trasformazione del frontend come servizio), rispetto alle quali il contributo intende presentare la cornice metodologica e i primi risultati ottenuti.

2. VOCI DI AUTORITÀ: ENTITÀ IN RELAZIONE

Nel modello logico/semantico della struttura dati ereditata dalla gestione del catalogo bibliografico, le voci di autorità rappresentavano entità con funzione di aggregazione e disambiguazione. L'eterogeneità delle forme che possono manifestare il riferimento a un'entità nominata veniva cioè ricondotta a una rappresentazione univoca declinata in molteplici forme varianti.
Di fatto cioè, all'interno di una base dati strutturata su un modello logico relazionale, le voci di autorità istituivano, in embrione, connessioni semantiche tra le risorse in catalogo [14, 1] relative, per esempio, a diverse forme di responsabilità, soggettazione, ecc.
Su questa base, è stata progettata l'estensione delle relazioni alle altre tipologie di risorse in catalogo, riducendo gli attributi delle entità descritte dalle voci di autorità a un set minimo, condiviso dai domini coinvolti, integrato dalla definizione di nuove tipologie di relazione e delle loro proprietà (vd. Fig. 1).

[2] Il software Geca è stato sviluppato dall'Istituto per la Matematica Applicata e Tecnologie Informatiche "E. Magenes" del CNR ed è oggetto di revisione, re-ingegnerizzazione e mantenimento da parte di un gruppo di lavoro dedicato, che ha progettato e sviluppato l'evolutiva sulla quale il presente contributo si innesta, precisandone le trasformazioni delle voci di autorità, del modello architetturale del front end e introducendo il modulo cartografico.

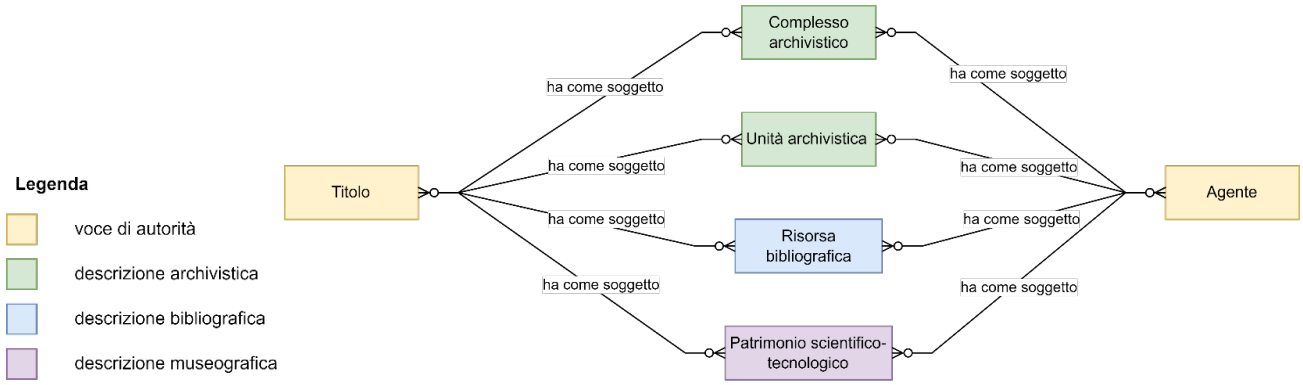


Figura 1. Diagramma ER che mostra le tipologie di relazioni instaurabili tra risorse e scheda Titoli e tra risorse e scheda Agenti

In questo modo, è stato possibile adattare il software alla specificità delle pratiche e metodi descrittivi, attivando però connessioni trasversali.
Un ruolo significativo hanno avuto, in questo senso, i luoghi, per i quali è stato proposto un tracciato descrittivo minimo all'interno del sistema, permettendo però di relazionare le entità con sistemi di rappresentazione esterni, come Geonames, specificamente dedicati alla modellazione di dati spaziali.
Il software ha così potuto identificare, a partire dalla creazione di relazioni *intra*-dominio (es. ha luogo di pubblicazione, stampa, conservazione, ecc.), connessioni spaziali (prossimità/distanza, inclusione/esclusione, ecc.) che hanno generato nuovi contenuti informativi e, conseguentemente, nuove possibilità di esplorazione *inter*-dominio.

3. CARTOGRAFIA DEL CATALOGO: ITINERARI TRA LE RELAZIONI

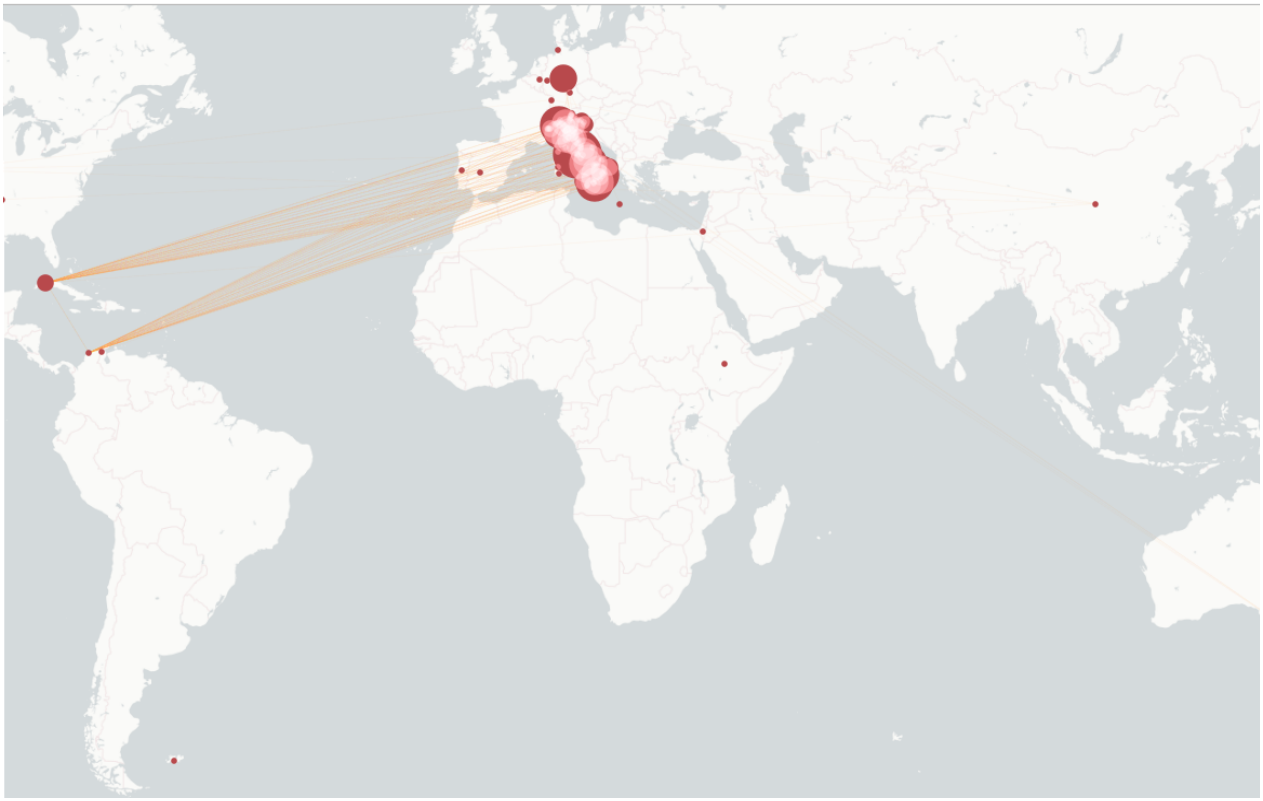


Figura 2. La mappa interattiva visualizza le risorse in catalogo, localizzando i luoghi (dimensionati per numero di risorse collegate) e generando relazioni spaziali a partire dalla relazione con le risorse (le linee collegano i luoghi relazionati con una stessa risorsa) [fonte base mappa: Carto].

La rappresentazione cartografica ha permesso di esplicitare queste relazioni a livello di presentazione del contenuto informativo. L'interfaccia utente è stata sviluppata integrando un modulo di visualizzazione dei dati spaziali attraverso una

mappa interattiva, sia come modalità di visualizzazione alternativa dei risultati di ricerca, sia a integrazione della visualizzazione delle schede di dettaglio, sia come modalità di navigazione integrata all'interno dell'intero catalogo (vd. Fig. 2).
La correlazione spaziale dei dati, determinata dalla prossimità della produzione o conservazione degli oggetti descritti, o dei riferimenti a luoghi citati contenuti nelle descrizioni stesse, veicola un contenuto informativo trasversale rispetto alle risorse in catalogo, capace di localizzarle nello spazio cartografico come contesto semantico-interpretativo astratto dal particolarismo informativo delle descrizioni disciplinari.
Il reticolo geografico, i confini amministrativi, lo spazio fisico ridotto alla rappresentazione simbolica della mappa attivano una dimensione informativa latente nell'indicizzazione catalografica, generando un nuovo livello ermeneutico della risorsa *per se*, ma attivando anche una diversa prospettiva sulla contestualizzazione della risorsa nel catalogo.
Le modalità di accesso ai dati e al loro contenuto informativo prevedevano già, per il software oggetto di sviluppo, modalità esplorative aggiuntive rispetto alla semplice ricerca per parametri o stringhe di ricerca [9]. Si prevedeva infatti la possibilità di creare contenitori narrativi, tipicamente costruiti su base tematica, utili a raggruppare le risorse in percorsi esplorativi orientati.
Questa modalità di presentazione è stata integrata dalla creazione di itinerari di collegamento tra i luoghi indicizzati, permettendo così all'utente non solo di osservare le relazioni spaziali tra le risorse, ma anche di crearne.
Il modulo cartografico si propone inoltre come base per l'integrazione della recente estensione dello standard IIIF per l'annotazione spaziale degli oggetti digitali, permettendo sia la georeferenziazione delle cartografie digitalizzate, sia la localizzazione su mappa delle digitalizzazioni allegate alle risorse in catalogo[3].

4. INTERFACCIA: LUOGO DI INCONTRO

Il modulo cartografico si innesta in una più generale trasformazione del modello architetturale strutturante il livello di presentazione del software, riprogettato come sistema *headless*. Il disaccoppiamento dell'interfaccia utente dalla *business logic*, attraverso lo sviluppo di uno strato intermedio di API che, garantendo la stabilità nell'esposizione dei dati, permette di ottenere maggiore flessibilità nelle soluzioni di frontend, ha aperto a diverse possibilità di sperimentazione.
L'obiettivo principale, all'origine di questa scelta progettuale, consiste nel tentativo di garantire un equilibrio tra il rigore scientifico-disciplinare che caratterizza la struttura dati e la varietà dei contesti comunicativi nei quali il software può essere impiegato, con la necessità di adattarsi a diversi requisiti di *user experience* [10].
Il primo sviluppo proposto è una soluzione basata sul framework opensource Wordpress, selezionato come piattaforma *low code* per la creazione di contenuti web. In particolare, a partire dagli strumenti e metodi condivisi all'interno del progetto Designers Italia (promosso da AgID), è stato elaborato un prototipo di tema per l'esplorazione di cataloghi di beni culturali (vd. Fig. 3). Il tema è strutturato per interfacciarsi direttamente con l'applicativo e garantire, senza necessità di sviluppo software, la personalizzazione di alcuni elementi di interfaccia.

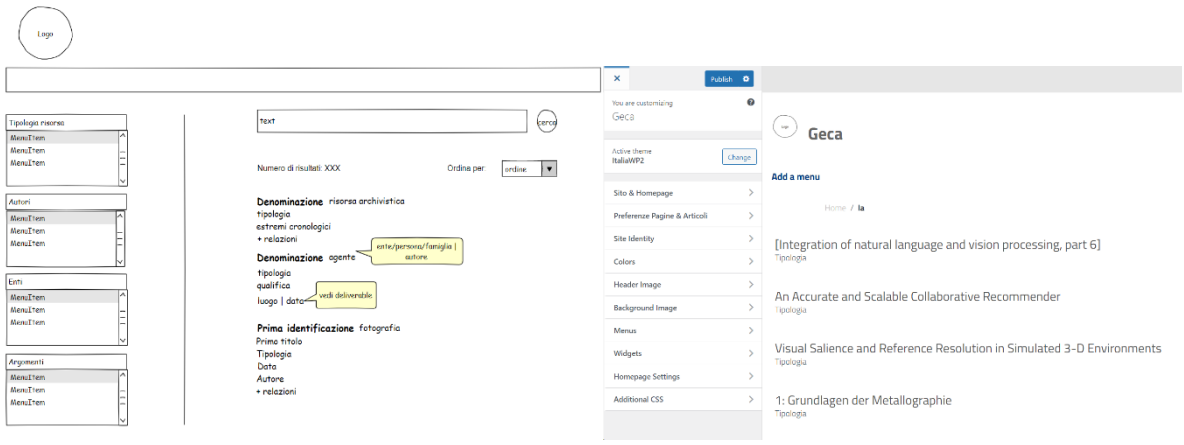


Figura 3. Il design di mockup a bassa fedeltà (a sinistra) è stato un processo partecipativo, che ha coinvolto gli esperti di dominio, finalizzato a definire gli elementi fissi e le possibilità di personalizzazione previste nello sviluppo del tema wordpress (a destra).

[3] https://iiif.io/api/extension/georef/

5. CONCLUSIONI

La progettazione e sviluppo di un sistema informativo è necessariamente un processo aperto, le cui linee evolutive sono sollecitate dal movimento degli orizzonti tecnologici, che lasciano intravedere prospettive di costante efficientamento delle prestazioni e funzionalità offerte dai software. Tuttavia, per quanto dinamica e flessibile, ogni soluzione architetturale impone vincoli che, garantendo robustezza, integrità, stabilità al sistema, non possono essere trascesi, riconducendo le trasformazioni possibili a un insieme finito, oltre il quale si rende necessaria una revisione strutturale e radicale.

L'evoluzione del software Geca è paradigmatica in questo senso: la sua estensione dal mondo bibliografico a quello GLAM, nel rispetto della specificità scientifica e metodologica degli ambiti disciplinari interessati, ha richiesto non solo l'integrazione di nuove possibilità descrittive, allargando lo spettro dei dati processabili dal sistema, ma anche l'attivazione di relazioni che convertissero questa eterogeneità in arricchimento informativo.

Il riconoscimento delle voci di autorità come nodi nevralgici per l'attivazione di questa rete di inferenze ha rappresentato un passaggio cruciale, i cui esiti potenziali e previsti non si limitano alla modifica del modello dati che organizza l'informazione a esse riferita, ma nel conferirle una natura semantica che tende a svincolarsi dal tracciato descrittivo, apre a una revisione più radicale della struttura della base dati nel suo complesso.

L'introduzione dell'informazione spaziale, il luogo, ha rappresentato un processo significativo per la relazione inversa tra il livello (generale) di granularità offerto dal tracciato descrittivo e la capillarità delle relazioni attivate. Infatti, la scelta di supportare il processo di entificazione attraverso la connessione con altri sistemi di rappresentazione presenti in rete, procede in una direzione di apertura del sistema verso l'esterno, interoperabilità e riuso dell'informazione esistente.

In corso di sviluppo, nell'ottica di implementare il ruolo dell'informazione spaziale come generatore di conoscenza inferita, è invece l'integrazione di strumenti di annotazione spaziale (come appunto incoraggiato da IIIF) e di estrazione di toponimi da campi testuali non strutturati.

La spazializzazione del dato però, come detto, non riguarda solo l'acquisizione, elaborazione e organizzazione dei dati, ma anche la loro presentazione e trasformazione in contenuto informativo. In quest'ottica, la specificità di Geca nell'offrire funzionalità di esplorazione dei contenuti in catalogo destinata anche a un pubblico di non esperti, incoraggia l'implementazione di forme di *storytelling* digitale che integrino l'esplorazione dello spazio fisico con l'esplorazione di contenuti aumentati (letteratura locative).

Prospettive di sviluppo accomunate da un'idea di base, manifestamente espressa in una soluzione architetturale che rimuove il vincolo della centralizzazione del livello di presentazione del sistema (front-end), per ridurlo (o elevarlo?) al rango di servizio, oggetto di progettazione partecipata, flessibile sia rispetto a molteplici contesti di fruizione, sia rispetto alla rapida evoluzione dei dispositivi e delle tecnologie che ne mediano l'accesso.

Non più, o non solo, luogo comune, l'interfaccia, cardo e decumano digitale che orienti cartesianamente l'esplorazione dei dati navigando gli utenti oltre l'orizzonte del *digital divide* che separa l'uomo dalla macchina, ma luogo *in* comune, tra l'uomo e la macchina, manifestazione della loro interazione, stato più che natura, determinato dalle variabili che compongono il contesto di utilizzo.

BIBLIOGRAFIA

[1] Biagetti, Maria Teresa, (a cura di). *Le ontologie bibliografiche: modelli concettuali e vocabolari condivisi per l'universo bibliografico.* Roma: Bulzoni, 2022.

[2] Bidney, Marcy, e Nathan Piekielek. «Towards a New Paradigm in Map and Spatial Information Librarianship». *J. Map Geogr. Libr* 14, fasc. 2–3 (2018): 67–74. https://doi.org/10.1080/15420353.2019.1662673.

[3] Bruni, Silvia, Francesca Capetta, Anna Lucarelli, Maria Gazia Pepe, Anna Peruginelli, e Marco Rulent. «Verso l'integrazione tra archivi, biblioteche e musei. Alcune riflessioni». *JLIS.it* 7, fasc. 1 (2016): 225–44. https://doi.org/10.4403/jlis.it-11482.

[4] Gold, Nicolas. «Service-Oriented Software in the Humanities: A Software Engineering Perspective». *Digital Humanities Quarterly* 3, fasc. 4 (2009).

[5] Hawkins, Andrew M. «Geographical Information Systems (GIS): their use as decision support tools in public libraries and the integration of GIS with other computer technology». *New Libr. World*, fasc. 1117 (1994): 4–13.

[6] Maggi, Roberta, Tiziana Pasciuto, Martina Mazzoleni, Maria Teresa Artese, Isabella Gagliardi, e Riccardo Albertoni. «GECA 3.0 - A new tool for cataloguing and enjoying cultural heritage». In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 373–79, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[7] McGlamery, Patrick, e Melissa Lamont. «Geographic information systems in libraries». *Database* 17, fasc. 6 (dicembre 1994): 35–44.

[8] Nichols, Gina L. «Merging Special Collections with GIS Technology to Enhance the User Experience». *Int. J. Stud. Res.* 5, fasc. 2 (2016). https://doi.org/10.31979/2575-2499.050205.

[9] Pasciuto, Tiziana, Riccardo Albertoni, Roberta Maggi, Maria Teresa Artese, Isabella Gagliardi, e Maurizio Gentilini. «Travelling Culture: Define, Implement, Enrich and Disseminate the Digital Cultural Heritage. The "DigitXL Project" Case Study». In *EDEN Research Workshop, Dubrovnik, 19-20/09/2022. Proceedings*, a cura di Josep M. Duart e Elena Trepule, 134–39. Dubrovnik, 2022.

[10] Porter, Catherine, Zenobie Garrett, Rebecca Milligan, Caleb Derven, e Ken Bergin. «Placing the archive online: a WebGIS approach to sharing library collections». *e-Perimetron* 17, fasc. 4 (2022): 161–80.

[11] Schreibman, Susan, Raymond G. Siemens, e John Unsworth, (a cura di). *A new companion to digital humanities*. Chichester: John Wiley & Sons Inc, 2016.

[12] Schreur, Philip E. «The Use of Linked Data and Artificial Intelligence as Key Elements in the Transformation of Technical Services». *Cataloging & Classification Quarterly* 58, fasc. 5 (2020): 473–85. https://doi.org/10.1080/01639374.2020.1772434.

[13] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[14] Zou, Xiaozhu, Siyi Xiong, Zhi Li, e Ping Jiang. «Constructing Metadata Schema of Scientific and Technical Report Based on FRBR». *Computer and Information Science* 11, fasc. 2 (2018). https://doi.org/10.5539/cis.v11n2p34.

Messaggistica Istantanea e Archivi Digitali. Quali soluzioni? Best practices e considerazioni dal contesto internazionale

Alessia Del Bianco
Università di Bologna, Italia - alessia.delbianco9@unibo.it

ABSTRACT

Negli ultimi anni la messaggistica istantanea – *instant messaging* (IM) – è diventata uno dei mezzi di comunicazione più utilizzati della nostra società: WhatsApp, Signal, Telegram, WeChat, ma anche Microsoft Teams o Google Chat, solo per citarne alcuni, sono diventati straordinariamente diffusi come metodo di comunicazione veloce ed efficiente. Come osserva Jenny Mitcham in un recente post dal titolo *What's up with using WhatsApp?* (*Digital Preservation Coalition*),[1] chi si occupa di archivi digitali dovrebbe iniziare ad interrogarsi sulla messaggistica istantanea come documento archivistico e pensare a come gestirla, archiviarla e conservarla a lungo termine, proprio in virtù del suo crescente impiego per comunicazioni importanti, aspetti decisionali o come parte delle nostre memorie digitali personali. Sebbene queste applicazioni favoriscano un modo di comunicare rapido e semplice, il loro utilizzo espone ad alcune criticità emergenti. L'intervento propone di delineare lo stato dell'arte sulla gestione della messaggistica istantanea negli archivi digitali: recenti studi e inchieste, svolti principalmente da istituti internazionali e archivi nazionali tra il 2020 e il 2023, hanno portano alla luce una prima riflessione attorno al concetto di messaggistica istantanea come *record,* nonché policy e best practices per la sua gestione e conservazione.

PAROLE CHIAVE

Messaggistica istantanea (IM); archivi digitali; digital preservation; best-practices.

1. INTRODUZIONE

Nel nostro quotidiano inviamo, e riceviamo, messaggi ad amici, familiari e sempre più spesso li utilizziamo nel contesto lavorativo, complice anche la pandemia di Covid-19 che ha influito sul nostro modo di relazionarci e comunicare. In anni relativamente recenti il loro impiego si è ampiamente diffuso tra il personale delle istituzioni governative – è recente la notizia riguardo la volontà del governo francese di adottare *Olvid* come piattaforma di messaggistica istantanea –,[2] o si pensi ai Comuni italiani che hanno adottato chat e canali WhatsApp o Telegram per comunicare con i cittadini. Non solo, queste applicazioni vengono largamente utilizzate per diffondere o condividere informazioni, materiale video, audio e documenti tra colleghi, imprese e altri professionisti.

È proprio dalle istituzioni governative e dagli archivi nazionali che arrivano le prime riflessioni sulla necessità di considerare l'*instant messaging* come documento archivistico e su come gestirlo di conseguenza. I messaggi, inviati o ricevuti, sono tendenzialmente considerati come un mezzo di comunicazione informale e, nella maggior parte dei casi, tendono a rivelarsi di natura transitoria, effimera e con un valore a breve termine – *short-lived* – ,[3] di conseguenza non necessariamente soggetti ad archiviazione. Per tale motivo, le istituzioni tendono a non adottare politiche per la gestione e conservazione, o ad averle per periodi di tempo molto limitati. Tuttavia, in un numero sempre crescente di casi, come ad esempio la *Covid Inquiry* che vede coinvolto Boris Johnson, l'informazione veicolata attraverso questi canali può essere prodotta nello svolgimento delle proprie attività istituzionali, fornendo prova e dando seguito ad aspetti decisionali e amministrativi rilevanti.[4] Tralasciando gli eventi, appare necessaria una riflessione sul contenuto della messaggistica istantanea come documento pubblico, come *record,* nonché sull'opportunità che questo debba essere registrato nel sistema

[1] Jenny Mitcham, *What's up with using WhatsApp?* In Digital Preservation Coalition, 31 March 2022; https://www.dpconline.org/blog/what-s-up-with-using-whatsapp

[2] Guillaume Grallet, *Olvid : qu'est-ce que cette messagerie instantanée dorénavant utilisée par le gouvernement?,* Le Point, 30 Novembre 2023; https://www.lepoint.fr/high-tech-internet/olvid-qu-est-ce-que-cette-messagerie-instantanee-dorenavant-utilisee-par-le-gouvernement-30-11-2023-2545205_47.php.

[3] C*omments of the Commission on a request for information from the European Ombudsman* - Strategic initiative SI/4/2021/TE; https://www.ombudsman.europa.eu/en/doc/correspondence/en/158616

[4] Peter Walker, *Boris Johnson faces tough questions at Covid inquiry over handling of pandemic Former prime minister to give evidence as mystery deepens over retrieval of WhatsApp messages*, The Guardian, December 6, 2023; https://www.theguardian.com/politics/2023/dec/05/many-of-boris-johnsons-whatsapps-cannot-be-retrieved-for-covid-inquiry

di gestione documentale e conservato a lungo termine per la nostra futura memoria. Richard Ovenden si è più volte pronunciato in merito, esprimendo preoccupazione riguardo l'utilizzo dei servizi di messaggistica nel governo britannico, in particolar modo verso le applicazioni che permettono di cancellare i messaggi automaticamente.[5]

Se in Italia il problema non è ancora stato trattato compiutamente [1],[6] l'analisi di direttive, inchieste e raccomandazioni, approcci teorici e pratici svolti nell'ambito internazionale da governi, istituti e archivi nazionali tra il 2020 e il 2023, ha permesso di tracciare uno stato dell'arte e di delineare le osservazioni che hanno condotto a una riflessione sulla messaggistica istantanea come *record*, nonché sulle possibili strategie di archiviazione. I report prodotti nei contesti governativi hanno permesso di individuare quale uso viene fatto dalla messaggistica istantanea nelle amministrazioni centrali, per quali comunicazioni, su quali applicazioni e dispositivi, se sono presenti pratiche condivise per la registrazione nei sistemi. A questo proposito sono rilevanti le inchieste dell'*Institut for Government for Government, WhatsApp in government. How ministers and officials should use messaging apps - and how they shouldn't* [2],[7] la *Strategic Initiative*[8] dell'Ombudsman europeo Emily O'Reilly,[9] così come vale la pena di segnalare il report *Gone in an Instant: How Instant Messaging Threatens the Freedom of Information Act*[10], dell' Cause of Action Institute (CoA Institute) and Americans for Prosperity Foundation (AFPF).[11] Best practices e strategie da seguire provengono altresì da archivi nazionali quali il *Nationaal Archief* Nederland (NA),[12] *National Archive of Australia* (NAA),[13] *Archive New Zealand*[14] o la *Directive Instant Messaging* del *Government of Newfoundland and Labrador,*[15] fino al bollettino del 5 febbraio 2023 del *National Archive Record Administration* (NARA)[16], nel quale si propone di estendere l'approccio *Capstone* alla messaggistica istantanea.[17]

Le indagini e le raccomandazioni portano alla luce un panorama non ancora pienamente definito sotto gli aspetti metodologici, applicativi e normativi. In particolare costituiscono tutt'ora ambiti di criticità condivisi sia le modalità di utilizzo, di trasparenza amministrativa e di sicurezza, sia quelli attinenti all'uso del patrimonio informativo. D'altra parte, la mancanza di un dibattito in merito, proprio in virtù del crescente utilizzo della messaggistica per comunicazioni importanti, pone diverse questioni da affrontare quali cosa può essere considerato come documento, policy e best practices per la corretta gestione del contenuto, individuando soluzioni condivise, possibili modelli ed eventuali proposte per la conservazione a lungo termine.

---

[5] Richard Ovenden, *Undelete our government*, in Digital Preservation Coalition, October 16, 2020; https://www.dpconline.org/blog/undelete-our-government

[6] Per una prima disamina della questione si veda Stefano Allegrezza*, "Effimeri ma non troppo". La conservazione dei messaggi istantanei: è ora di cominciare ad occuparcene?* in AIDAinformazioni, no. 1-2, gennaio-giugno 2023, Bari: Caucci editore, 2023, pp. 9-32; si veda inoltre ParER, Polo Archivistico Regione Emilia Romagna. *Dallo scandalo dei lockdown files un nuovo monito sull'importanza della conservazione digitale*, 19 dicembre 2023; https://poloarchivistico.regione.emilia-romagna.it/news-in-evidenza/dallo-scandalo-dei-lockdown-files-un-nuovo-urgente-monito-sullimportanza-fondamentale-della-conservazione-digitale

[7] Tim Dourrant, Alice Lilly, Paeony Tingay, *WhatsApp in government. How ministers and officials should use messaging apps – and how they shouldn't*, Institute for Government, London 2022; https://www.instituteforgovernment.org.uk/publication/whatsapp-government

[8] O'Reilly, Emily (European Ombudsman). *Closing note on the strategic initiative on how EU institutions, bodies, offices and agencies record text and instant messages sent/received by staff members in their professional capacity (Case SI/4/2021/MIG)*, Strasbourg, 13/07/2022. https://www.ombudsman.europa.eu/it/doc/correspondence/it/158383

[9] O'Reilly, Emily. *Closing note on the strategic initiative on how EU institutions, bodies, offices and agencies record text and instant messages sent/received by staff members in their professional capacity* (SI/4/2021/MIG), Strasbourg, 13/07/2022; https://www.ombudsman.europa.eu/en/doc/correspondence/en/158383

[10] Kimbrell, Thomas; Valvo, James; Schmidt, Kevin. *Gone in an Instant: How Instant Messaging Threatens the Freedom of Information Act,* Cause of Action Institute, Americans for Prosperity Foundation, 2020. https://causeofaction.org/gone-in-an-instant-how-instant-messaging-threatens-the-freedom-of-information-act/

[11] Thomas Kimbrell, James Valvo, Kevin Schmidt, *Gone in an Instant: How Instant Messaging Threatens the Freedom of Information Act*, Cause of Action Institute, Americans for Prosperity Foundation, 2020; https://causeofaction.org/gone-in-an-instant-how-instant-messaging-threatens-the-freedom-of-information-act/

[12] Nationaal Archif Nederland, *Informatieblad archiveren chatberichten. Over de mogelijkheden voor duurzame toegankelijkheid van chatberichten*, 2019; https://www.nationaalarchief.nl/archiveren/kennisbank/informatieblad-archiveren-chatberichten

[13] National Archive of Australia, *Managing social media and instant messaging (IM),* https://www.archives.govt.nz/manage-information/how-to-manage-your-information/digital/text-messages-and-other-communications

[14] New Zealand Archives, *Text messages and other communications*; https://www.archives.govt.nz/manage-information/how-to-manage-your-information/digital/text-messages-and-other-communications

[15] Government of Newfoundland and Labrador, Office of the Chief Information Officer, *Directive Instant Messaging*, OCIO Reference: DOC00967/2012, 2020; https://www.gov.nl.ca/exec/ocio/im/policy-instruments/instant-messaging-to-remove/

[16] Steidel Wall, Debra. *Expanding the Use of a Role-Based Approach (Capstone) for Electronic Messages*, The U.S. National Archives and Records Administration, NARA Bulletin 2023-02, January 5, 2023. https://www.archives.gov/records-mgmt/bulletins/2023/2023-02

[17] Debra Steidel Wall, *Expanding the Use of a Role-Based Approach (Capstone) for Electronic Messages*, NARA Bulletin 2023-02, January 5, 2023; https://www.archives.gov/records-mgmt/bulletins/2023/2023-02

2. INSTANT MESSAGING AS A RECORD

Dalle inchieste e raccomandazioni emergono alcune riflessioni significative riguardo l'*instant messaging* come *record.* In merito si è espresso l'Ombudsman europeo Emily O'Reilly nel 2021. Il Mediatore ha avviato una *Strategic Initiative* che ha l'obiettivo di fare il punto su come le istituzioni, gli organi, gli uffici e le agenzie dell'Unione Europea registrano i messaggi di testo e istantanei inviati/ricevuti dal personale nella loro veste professionale, *How EU institutions, bodies, offices and agencies record text and instant messages sent/received by staff members in their professional capacity*[18]. L'Ombudsman ha chiesto a tutte le agenzie (European Parliament; Council of the European Union; European Border and Coast Guard Agency; European Chemicals Agency; European Food Safety Authority; European Medicines Agency; European Commission; European Central Bank) chiarimenti sulle norme adottate in materia, sulle modalità di registrazione e sulla presenza di linee guida e istruzioni per il personale. Il fine ultimo è quello di redigere un elenco di buone pratiche per aiutare le amministrazioni. Dall'indagine emergono importanti considerazioni riguardo regole e pratiche impiegate, o meno, nonché sull'importanza di considerare i messaggi come *record.* Per il Mediatore i messaggi di testo e istantanei rientrano nell'ambito di applicazione della legge dell'EU sull'accesso pubblico ai documenti e «Concretely, this implies that the decision to record a certain piece of informaton in the administraton's document management system should not be dependent on the medium - be it a letter, an email, a text or instant message - but on its content»[19].

Ancor prima dell'indagine europea, la gestione delle informazioni scambiate tramite le applicazioni di messaggistica è diventata attuale nel contesto neerlandese tra il 2019 e il 2020, a seguito della sentenza del Consiglio di Stato del 20 marzo 2019: *WhatsApp en SMS-berichten op zowel zakelijke als privételefoons van bestuurders en ambtenaren vallen onder de Wet openbaarheid van bestuur (Wob), als deze in het kader van het werk zijn verstuurd*[20], WhatsApp e i messaggi di testo su telefoni aziendali e privati di dirigenti e funzionari pubblici rientrano nella legge sull'accesso pubblico se sono stati inviati nell'ambito del lavoro. Una successiva sentenza del Consiglio di Stato del 21 ottobre 2020 conferma che SMS e messaggi WhatsApp utilizzati nello svolgimento di compiti governativi rientrano nel campo di applicazione dell'allora Wob (da maggio 2022 sostituito dal *Woo-Open Government Act*).[21]

Sono altrettanto significative le dichiarazioni del *National Archives and Records Administration*: alla domanda *Does IM Content Qualify as a Federal Record?* afferma «Agencies that allow IM traffic on their networks must recognize that such content may be a Federal record under that definition and must manage the records accordingly», ribadendo nel più recente bollettino del febbraio 2023[22] che i messaggi elettronici «created or received in the course of agency business are likely federal records. This includes electronic messages sent or received on personal devices that meet the definition of a record. These messages must be forwarded or copied to an official account within 20 days».[23] Anche altri archivi nazionali, come il *National Archive of Australia* o *Archives New Zealand* specificano che messaggi creati o ricevuti nel corso delle attività di governo sono da considerarsi *records,* «If an organisation uses text messaging or any other instantaneous, non-sequential electronic communication mechanism to conduct business, e.g. social media, these communications are considered records under the Public Records Act 2005. As such, they must be managed accordingly».[24]

WhatsApp in Government [2], a seguito di un'indagine che rileva l'utilizzo e la gestione delle app di messagistica nel governo britannico, mette in evidenza come l'impiego di WhatsApp stia trasformando il modo in cui i politici, e altre personalità, discutono e prendono decisioni. Cosa si deve fare? Per l'istituto bisogna chiarire le pratiche di *record-keeping* e quali messaggi devono essere conservati. L'uso di questi canali di comunicazione, per le attività ufficiali, rende difficile

[18] O'Reilly, Emily (European Ombudsman). *How EU institutions, bodies, offices and agencies record text and instant messages sent/received by staff members in their professional capacity* (Case SI/4/2021/MIG), Strasbourg, 30/06/2021. https://www.ombudsman.europa.eu/it/doc/correspondence/en/143787

[19] O'Reilly, Emily *The recording of text and instant messages sent/received by staff members in their professional capacity - Practical recommendations for the EU administration*, (SI/4/2021/MIG), Strasbourg, 14/07/2022. https://www.ombudsman.europa.eu/en/doc/correspondence/en/158302; https://www.ombudsman.europa.eu/en/doc/correspondence/en/158383

[20] Raad Van State (NL). Sms'jesWhatsApp-berichten op zakelijke én privételefoons zijn te 'wobben' ECLI:NL:RVS:2019:899, 20/03/2019. https://www.raadvanstate.nl/actueel/nieuws/@114494/sms-jes-whatsapp/

[21] Raad Van State. *Sms'jesWhatsApp-berichten op zakelijke én privételefoons zijn te 'wobben'*, 20 maart 2019. https://www.raadvanstate.nl/@114494/sms-jes-whatsapp/; cfr. Afdeling bestuursrechtspraak van de Raad van State, ABRvS 20 maart 2019, ECLI:NL:RVS:2019:899; Afdeling bestuursrechtspraak van de Raad van State, ABRvS 21 oktober 2020, ECLI:NL:RVS:2020:2477.

[22] Steidel Wall, Debra. *Expanding the Use of a Role-Based Approach (Capstone) for Electronic Messages*, The U.S. National Archives and Records Administration, NARA Bulletin 2023-02, January 5, 2023. https://www.archives.gov/records-mgmt/bulletins/2023/2023-02

[23] https://www.archives.gov/records-mgmt/bulletins/2023/2023-02

[24] https://www.archives.govt.nz/manage-information/how-to-manage-your-information/digital/text-messages-and-other-communications

la gestione dei documenti ed è fondamentale che il governo adotti delle soluzioni, anche tecnologiche, per garantire che i messaggi importanti vengano conservati a lungo termine per indagini, per ricorsi in tribunale, per gli archivi nazionali e per la memoria istituzionale, nonché per trasparenza amministrativa.[25]

A mettere in luce le criticità di una mancata archiviazione dei messaggi è il report *Gone in an Instant: How Instant Messaging Threatens the Freedom of Information Act*[26], della Cause of Action Institute and Americans for Prosperity Foundation. La relazione riferisce come numerose agenzie violino la legge sui documenti federali, e le indicazioni dell'archivio nazionale, non mantenendo i messaggi istantanei. La mancata conservazione da parte delle agenzie dei documenti creati attraverso le piattaforme di messaggistica, che sono prevalenti sul posto di lavoro, rischia di compromettere il *Freedom of Information Act* (FOIA).[27]

Solo per riportare un ultimo esempio, le policy sull'uso della posta elettronica e della messaggistica istantanea della University of Waterloo (Canada) sottolineano la necessità di considerare le e-mail e i messaggi istantanei, creati e ricevuti dai dipendenti nel corso del loro lavoro, come documenti dell'università, nella stessa misura in cui lo sono le informazioni contenute in altri supporti, come i documenti elettronici e cartacei.[28]

3. BEST PRACTICES PER LA GESTIONE E CONSERVAZIONE

La messaggistica istantanea crea una serie di sfide dal punto di vista della gestione degli archivi digitali, tra cui la ricerca dei documenti che rispondono alle richieste di accesso, l'esigenza che siano conservati e custoditi secondo le corrette pratiche, l'acquisizione nel sistema di gestione, il mantenimento dei contenuti, dei dati contestuali e la resa comprensibile e utile.

Le istituzioni che hanno avviato le già sopracitate indagini, riconoscendo la necessità di trattare la messaggistica istantanea come *record*, hanno formulato delle raccomandazioni e best practices riguardo l'utilizzo, la gestione e la conservazione. Il NARA[29] nel bollettino del febbraio 2023 comunica la volontà di ampliare il metodo *Capstone* – un sistema basato sui ruoli – alla messagistica istantanea. Per il Federal Record Management è «un approccio di successo per la gestione delle e-mail, dei messaggi di testo e di altre comunicazioni digitali, nell'ambito di un tentativo di aiutare le agenzie a gestire una mole sempre crescente di documenti elettronici». Le agenzie federali dovranno adottare processi e strumenti per la conservazione dei messaggi scambiati tramite chat o applicazioni di terze parti, da parte delle loro figure apicali e manageriali, per un tempo che va dai 3 ai 7 anni o dai 15 ai 30 anni, in base alla tipologia del contenuto. In particolare, i record di determinati ruoli, o posizioni, possono essere classificati come permanenti per il trasferimento all'archivio nazionale.[30]

L'*Institute for Government* [2], a seguito dell'inchiesta, rilascia alcune valide indicazioni, quali: i ministri, i consiglieri e i funzionari non dovrebbero usare i telefoni personali per le attività governative di rilievo; i dipartimenti devono assicurarsi che i messaggi WhatsApp rilevanti siano conservati a lungo termine; i dipartimenti devono assicurarsi che WhatsApp non ostacoli la trasparenza o le responsabilità. Non ultima, la raccomandazione di assicurarsi che i messaggi vengano trasferiti al sistema di gestione e successivamente all'archivio nazionale dopo vent'anni. Non è un caso la nota del *Cabinet Office* (UK) del marzo 2023 in cui si concordano nuove direttive riguardo all'utilizzo di *Non-Corporate Communication Channels For Government Business. Using non-corporate communication channels (e.g. Whatsapp, private email, sms) for government business*, una guida che deve essere un riferimento per ministri, funzionari e altri dipendenti nel valutare con consapevolezza l'opportunità di utilizzare questi canali.[31]

L'Ombudsman europeo, al pari dell'*Institute for Government*, definisce delle prassi per gestire la messaggistica istantanea. In *The recording of text and instant messages sent/received by staff in their professional capacity*[32] si ritrovano una serie di raccomandazioni che possono guidare l'amministrazione dell'Unione Europea nella gestione dei messaggi di testo e

[25] https://www.instituteforgovernment.org.uk/publication/whatsapp-government

[26] Kimbrell, Thomas; Valvo, James; Schmidt, Kevin. *Gone in an Instant: How Instant Messaging Threatens the Freedom of Information Act*, Cause of Action Institute, Americans for Prosperity Foundation, 2020.

[27] https://causeofaction.org/gone-in-an-instant-how-instant-messaging-threatens-the-freedom-of-information-act/

[28] University of Waterloo, *E-mail & Instant Messaging. Guidelines on use of e-mail and instant messaging*; https://uwaterloo.ca/privacy/policies-guidelines/e-mail-instant-messaging

[29] Steidel Wall, Debra. *Expanding the Use of a Role-Based Approach (Capstone) for Electronic Messages*, The U.S. National Archives and Records Administration, NARA Bulletin 2023-02, January 5, 2023. https://www.archives.gov/records-mgmt/bulletins/2023/2023-02

[30] https://www.archives.gov/records-mgmt/bulletins/2023/2023-02

[31] Cabinet Office UK, *Non-Corporate Communication Channels for Government Business. Using non-corporate communication channels (e.g. Whatsapp, private email, sms) for government business*, 30 March 2023; https://www.gov.uk/government/publications/non-corporate-communication-channels-for-government-business

[32] O'Reilly, Emily (European Ombudsman). *The recording of text and instant messages sent/received by staff members in their professional capacity - Practical recommendations for the EU administration*, (Case SI/4/2021/MIG), Strasbourg, 14/07/2022. https://www.ombudsman.europa.eu/en/publication/en/158302

istantanei. Tra le varie avvertenze, si sottolinea l'esigenza di sensibilizzare e fornire al personale una guida chiara su come estrarre, trasferire e registrare i messaggi di testo e istantanei, procedure che dovrebbero essere svolte regolarmente. Si raccomanda di mettere in atto soluzioni tecnologiche che consentano di registrare facilmente i messaggi nei sistemi di gestione, nel frattempo prevedere modalità alternative. Assicurarsi inoltre che i periodi di conservazione dei messaggi contenuti nei dispositivi, utilizzati per lavoro, siano in linea con le politiche adottate.[33]

Una prassi comune a molte amministrazioni, e che si rivela strategica in assenza di soluzioni condivise e tecnologiche adeguate, è fornire ai propri dipendenti una serie di domande – pur non esaustive – che guidano il dipendente nel riconoscimento di un messaggio potenzialmente importante e che deve essere archiviato, indicando possibili tipologie di messaggi, come riconoscerli e quale procedura adottare.

Il *Nationaal Archief* (NL)*,* a seguito delle dichiarazioni del governo neerlandese, ha elaborato una guida per le organizzazioni governative attraverso l'*Informatieblad archiveren chatberichten. Over de mogelijkheden voor duurzame toegankelijkheid van chatberichten*.[34] Qui vengono fornite possibili soluzioni per l'archiviazione dei messaggi, suggerite indagini preliminari per individuare applicazioni e comunicazioni, nonché modalità per determinare una lista di selezione o soluzioni tecniche per l'export delle chat. Le istruzioni non si limitano a WhatsApp, o ai soli SMS, ma a tutti i messaggi di testo prodotti con qualsiasi applicazione, come Telegram, Signal, Instagram o Facebook Messenger: se utilizzati per svolgere compiti governativi diventano rilevanti.

Non solo NARA o il *Nationaal Archief,* anche l'*Archives New Zealand* e il *National Archive of Australia* offrono delle buone prassi da seguire,[35] avvalorando sempre più la necessità di provvedere a indicazioni pratiche che incontrano le esigenze delle organizzazioni, sia per l'utilizzo dei sistemi di messaggistica disponibili per svolgere le proprie attività, sia per una corretta gestione e conservazione.

4. CONCLUSIONI E PROSPETTIVE FUTURE

La formulazione di inchieste e raccomandazioni può essere un impulso a sviluppare una riflessione, ancora assente in ambito nazionale, sul tema. Nella pubblica amministrazione italiana, social media e applicazioni di messaggistica sono diffusi quali strumenti di supporto per comunicare con i cittadini e come servizio di pubblica utilità, ad esempio le chat *one to one* o i canali WhatsApp e Telegram promossi da alcuni comuni italiani.[36] Non solo, anche le piattaforme di collaborazione come Microsoft Teams sono sempre più utilizzate dal personale nel contesto lavorativo. Tuttavia, si opera ancora in un campo non normato, pur incoraggiato dalle ultime riforme come le recenti indicazioni sull'utilizzo dei social media da parte del Codice di Comportamento d.P.R. n. 81/2023 e l'invito ad adottare delle Social Media Policy interne ed esterne, in cui si ritrovano alcuni riferimenti all'utilizzo delle applicazioni di messaggistica.[37]

La gestione dei messaggi negli archivi digitali è un tema emergente. Le raccomandazioni del Mediatore europeo non hanno ancora avuto delle ricadute nell'ambito nazionale, manca un inquadramento generale sull'uso della messaggistica istantanea nella pubblica amministrazione italiana, ed è difficile riscontrare politiche e pratiche condivise da parte della comunità archivistica. Un monito sull'importanza della conservazione digitale arriva dalla newsletter del ParER (Polo archivistico dell'Emilia Romagna) riguardo lo scandalo dei *lockdown files.* Come sottolineato nell'articolo, i *lockdown files* hanno riportato l'attenzione sulla «riflessione su se e quanto i personaggi politici, o comunque ai massimi vertici dei nostri sistemi sociali, possano affidarsi all'uso di media personali, e per questo non coperti dalle normali regole di registrazione a archiviazione per finalità ufficiali e di documentazione storica, nell'esercizio delle proprie funzioni».[38]

Emergono, in conclusione, alcune criticità generali: la mancanza di linee guida unitarie - spesso ci si affida a raccomandazioni dei singoli istituti -, la fragilità dei messaggi, l'utilizzo dei propri dispositivi personali, il problema di individuare cosa è considerato come un *record* o meno, con il conseguente rischio di archiviare e conservare tutto, anche

[33] https://www.ombudsman.europa.eu/en/doc/correspondence/en/158302

[34] Nationaal Archif Nederland, *Informatieblad archiveren chatberichten. Over de mogelijkheden voor duurzame toegankelijkheid van chatberichten,* 2019; https://www.nationaalarchief.nl/archiveren/kennisbank/informatieblad-archiveren-chatberichten

[35] https://www.naa.gov.au/information-management/types-information-and-systems/types-information/managing-social-media-and-instant-messaging-im; https://www.archives.govt.nz/manage-information/how-to-manage-your-information/digital/text-messages-and-other-communications

[36] Si veda, ad esempio, Agenda Digitale, «Utilizzo di WhatsApp PA e comunicazione social: il Comune di Arezzo chatta con i cittadini su WhatsApp», https://www.agendadigitale.eu/cittadinanza-digitale/pa-e-comunicazione-social-il-comune-di-arezzo-chatta-con-i-cittadini-su-WhatsApp/ e https://www.comune.arezzo.it/unico-online-WhatsApp-policy

[37] Si vedano, solo per citarne alcune, le Social Media Policy dell'Azienda Ospedaliera Universitaria di Sassari, AOU di Sassari, Marzo 2021; http://www.aousassari.it/documenti/11_591_20210927155653.pdf; le Social Media Policy dell'Università Statale di Milano, https://www.unimi.it/it/ateneo/normative/policy/social-media-policy o il Codice di Comportamento dell'Azienda Ospedale Università Padova, https://www.aopd.veneto.it/index.cfm?action=mys.apridoc&iddoc=2526

[38] https://poloarchivistico.regione.emilia-romagna.it/news-in-evidenza/dallo-scandalo-dei-lockdown-files-un-nuovo-urgente-monito-sullimportanza-fondamentale-della-conservazione-digitale

il non necessario, o di tagliare fuori ciò che invece potrebbe essere fondamentale. D'altra parte, come evidenziato nelle raccomandazioni dell'*Archives New Zealand,* impedire semplicemente alle persone di adoperare la messaggistica elettronica e altri sistemi di comunicazione per svolgere attività ufficiali è sproporzionato, difficile da applicare e non tiene conto dei vari modi in cui i dipendenti comunicano.[39]

Quali sono le soluzioni possibili per la gestione della messaggistica nelle amministrazioni? Alcune proposte da cui si dovrebbe partire:

- Sviluppare una riflessione in merito alla messaggistica istantanea come documento archivistico.
- Individuare buone pratiche, procedure chiare e soluzioni tecnologiche condivise.
- Identificare i messaggi di testo rilevanti: identificare i messaggi di testo e le altre comunicazioni che hanno rilevanza a lungo termine, che sono transitorie o che hanno un valore a breve termine.
- Introdurre revisioni regolari: garantire che politiche e processi siano aggiornati con i cambiamenti tecnologici
- Formare il personale: assicurare che il personale sia formato con metodi semplici e pratici per gestire i messaggi di testo e altre comunicazioni.
- Configurare le tecnologie: consentire l'acquisizione di messaggi elettronici e relativi metadati.

L'assenza di azioni tempestive potrebbe portare a una scomparsa inevitabile di informazioni, privando potenzialmente storici e ricercatori di fonti importanti per la nostra storia.

BIBLIOGRAFIA

[1] Allegrezza, Stefano. «Effimeri ma non troppo. La conservazione dei messaggi istantanei: è ora di cominciare ad occuparcene?» *AIDAinformazioni*, fasc. 1–2 (2023): 9–32.

[2] Dourrant, Tim, Alice Lilly, e Paeony Tingay. «WhatsApp in government. How ministers and officials should use messaging apps – and how they shouldn't». London: Institute for Government, 2022.

[39] https://www.archives.govt.nz/manage-information/how-to-manage-your-information/digital/text-messages-and-other-communications

Nuove interazioni con collezioni digitali: l'"Archivio digitale del Capitolo di Laterza"

Stefania Riso[1], Nicola Barbuti[2]
[1] Università degli Studi di Bari Aldo Moro, Italia - stefania.riso@uniba.it
[2] Università degli Studi di Bari Aldo Moro, Italia - nicola.barbuti@uniba.it

ABSTRACT

Il paper presenta gli esiti di una ricerca finalizzata a rigenerare in ambiente digitale una serie documentale relativa a una plurisecolare vertenza sugli *Usi Civici*. Originariamente, la documentazione era conservata presso l'Archivio del Capitolo di Laterza, poi smembrata in seguito agli eventi storici e giuridici succedutisi nel corso di circa quattro secoli dal XVII al XX. La parte più consistente della serie è confluita nel tempo nell'Archivio privato della Famiglia dell'Aquila, mentre una porzione residuale di documenti è oggi conservata presso l'Archivio della Chiesa Matrice di S. Lorenzo Martire di Laterza. Il progetto ha previsto originariamente la digitizzazione di entrambi gli insiemi documentali e la ricomposizione in ambiente digitale della primigenia organizzazione archivistica dei documenti nell'antico Archivio Capitolare. In seguito, l'analisi diretta degli originali ha evidenziato alcune caratteristiche che hanno aperto una nuova prospettiva di fruizione della collezione digitale. Sono state rilevate fascicolazioni e cartulazioni relative ad aggregazioni di documenti funzionali a diverse fasi del contenzioso giuridico. Si è, inteso, dunque, rendere disponibili agli utenti le singole aggregazioni clusterizzandole per cartulazione di riferimento. Attualmente, si stanno studiando le possibili soluzioni di interfacce di interrogazione atte a facilitare l'accesso e la consultazione degli utenti.

PAROLE CHIAVE

Digital Heritage; digitizzazione; beni documentali; archivi digitali; open data.

1. INTRODUZIONE

Nel 2022 è stato pubblicato il Piano Nazionale di Digitalizzazione del patrimonio culturale (PND)[1], il cui obiettivo è promuovere e coordinare la digitizzazione e la pubblicazione in rete di 75 milioni di oggetti digitali rappresentativi di beni culturali entro il 31 dicembre 2026. Il Piano riconosce gli oggetti digitali nella loro funzione di espressioni e manifestazioni sociali e culturali della trasformazione digitale. La creazione di questa imponente massa critica di oggetti digitali a contenuto culturale apre la riflessione su quali siano i processi di digitizzazione e di creatività digitale, dai quali generare risorse il cui valore culturale sia determinato non in quanto *digital twins* dei beni rappresentati, ma per la loro funzione di espressioni e manifestazioni che, nel tempo, siano valorizzabili quali «testimoni e memoria storica da trasferire nell'integrità dei processi nel tempo e nello spazio» [4]. In tale direzione, alcune recenti esperienze di *musealizzazione phygital* [8], nelle quali sono state create soluzioni in prospettiva *user-oriented*, rappresentano validi precedenti da considerare, nell'ottica di programmare strategie di digitizzazione centrate sulle interazioni degli utenti [9] che emancipino dall'autoreferenzialità le iniziative realizzate o in corso. Alcune progettualità esemplificative di tale visione sono *Curious Alice: the VR experience*[2] e *Immersive Dickens*[3] del Victoria and Albert Museum di Londra, o ancora *The Quintana 4D Museater Lab (Q4D)* realizzato dal CRHACK Lab di Foligno [5].

In questo contributo si presentano i risultati intermedi di una ricerca e della relativa sperimentazione, finalizzate a definire un modello di rigenerazione in ambiente digitale di complessi archivistici fisicamente smembrati. Quale caso di studio è stata considerata una serie documentale anticamente conservata presso l'Archivio del Capitolo di Laterza, smembrata nel tempo in due insiemi a causa delle vicende storiche e giuridiche determinate da una plurisecolare vertenza sugli *Usi Civici*[4]. In prima istanza, obiettivo della ricerca era ricostituire in ambiente digitale l'originaria consistenza e organizzazione della serie, da tempo ormai non più consultabile nella sua omogeneità e integrità. In corso d'opera, l'analisi della documentazione originale ha rivelato fascicolazioni e cartulazioni di mani ed epoche diverse, riconducibili ad aggregazioni e riusi dei documenti distanti nel tempo, funzionali alle varie fasi del contenzioso giuridico. Ne è sorta l'idea di provare a rendere le varie aggregazioni accessibili agli utenti, ricostruendole e riproponendole in ambiente digitale, nella prospettiva di

[1] http://pnd.beniculturali.it
[2] https://www.vam.ac.uk/articles/curious-alice-the-vr-experience
[3] https://www.vam.ac.uk/research/projects/immersive-dickens
[4] Gli *Usi Civici* sono diritti di godimento collettivo su beni appartenenti al demanio, a un comune o a un privato, spettanti ai membri di una collettività. Questi diritti spesso derivano da antichi privilegi concessi dalla nobiltà o dallo Stato e sono finalizzati a garantire l'uso comune di risorse quali boschi, pascoli, terreni agricoli o risorse idriche.

intercettare le esigenze e l'interesse sia dei professionisti del settore archivistico, sia di altre comunità di utenti non necessariamente esperti di dominio.
Allo scopo, è stato ripreso e valutato un tracciato di metadati in formato open (vd. Figg. 1a-1b) già utilizzato per la digital library *Open Memory*[5], nell'ottica di integrarlo con elementi utili a indicizzare e gestire le diverse cartulazioni, di modo da generare le clusterizzazioni coerenti con ciascuna fase in risposta alle interrogazioni e interazioni degli utenti.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Identificativo directory	Identicativo risorsa originale	Posizione della risorsa analogica <N. fascicolo>	Estremi cronologici 1 <data cronica>	Estremi cronologici 2 <data topica>	Oggetto/ Soggetto del fascicolo	Descrizione	Soggetto Responsabile della risorsa analogica	Persona giuridica (Contributore 1)	Persona fisica (Contributore 2)	Altro (Contributore 3)	Consistenza del fascicolo	Note	Genere risorsa analogica
2														
3														
4														
5														
6														
7														
8														
9														
10														
11														
12														
13														
14														

Foglio 1

Figura 1a. Tracciato in formato XLSX - sezione Oggetto Analogico

	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS
1	Contenuto risorsa digitale	Livello di descrizione	Soggetto responsabile del progetto	Soggetto produttore 1	Soggetto produttore 2	Titolo del progetto	Consistenza risorsa digitale	Genere della risorsa	Formato pubblicata	Data e ora di creazione della risorsa digitale	Risoluzione spaziale	Dimensioni della risorsa digitale	Profilo ICC
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													

Foglio 1

Figura 1b. Tracciato in formato XLSX - sezione Oggetto Digitale

Attualmente, è in corso lo studio e l'implementazione di un'interfaccia utente che metta a disposizione funzionalità di interrogazione delle fonti puntando anche alle singole aggregazioni per cartulazione, oltre a quelle comunemente in uso.
Nell'ottica di rappresentare meglio quali siano stati gli approcci sfidanti valorizzati nella ricerca, si delinea di seguito lo stato dell'arte del complesso documentale originario.

2. L'ARCHIVIO DEL CAPITOLO COLLEGIALE DI LATERZA

La Chiesa Matrice di S. Lorenzo Martire di Laterza, costruita negli anni 1408-1414 [7: 163]; [11: 166] e consacrata il 19 novembre 1673, era retta dal Capitolo laertino il cui archivio era originariamente conservato presso la sagrestia. Nel 1908 il Capitolo si estinse con la morte dell'ultimo partecipante e il suo archivio confluì nell'Archivio della Chiesa di S. Lorenzo. Tra il 1978 e il 1987 la Chiesa di S. Lorenzo fu oggetto di un radicale restauro strutturale che ne provocò la chiusura al culto fino al 1983 [11: 168-170]. In questi anni, il timore che le carte d'archivio potessero correre rischi di dispersioni o danneggiamenti spinse la Sovrintendenza archivistica per la Puglia a intervenire sul complesso archivistico ecclesiastico,

[5] Si tratta di una Digital Library in Open Data progettata da D.A.BI.MUS. Srl, al fine di raccontare storie di interesse e consentire l'accesso, il riutilizzo e la ridistribuzione delle risorse digitali storiche, promuovendo nuove forme di partecipazione sociale e l'interazione attiva delle comunità antropiche con un nuovo Patrimonio digitale. Attualmente sono disponibili in Open Memory 31 dataset frutto di tre progetti di digitizzazione di archivi storici, avviati nel corso degli ultimi tre anni: l'Archivio Storico della Colonia Penale delle Isole Tremiti, l'Archivio storico della Fondazione Giuseppe e Salvatore Tatarella, e l'Archivio Storico della Chiesa di Santa Maria Matrice di Palo del Colle (in provincia di Bari). http://www.openmemory.eu/

dichiarandolo di notevole interesse storico, con declaratoria del 25 febbraio 1987, ai sensi del D.P.R. n. 1409/1963 e L. n. 253/1986, e avviando operazioni di riordino e inventariazione.
Il risultato delle attività svolte dai funzionari della Sovrintendenza fu la redazione dell'inventario dell'Archivio della Chiesa Matrice di S. Lorenzo Martire di Laterza [14], pubblicato nel 1993 in un volume curato da Carlo dell'Aquila, dal titolo "Per la Storia di Laterza. Fonti archivistiche e documentarie", e inserito nella "Biblioteca di Cultura Pugliese", diretta da Mario Congedo, con il n. 74 della seconda serie [12].
L'Archivio storico della Chiesa Matrice di S. Lorenzo Martire di Laterza risulta attualmente costituito da quattro sezioni [14: 23-55]:

- Fondo pergamenaceo, consistente in 25 pergamene datate tra il 1549 e il 1717
- Fondo cartaceo, consistente in 52 buste e formato essenzialmente da carteggi e registri
- Libri canonici, consistenti in 104 volumi
- Codici liturgici musicali, consistenti in 9 volumi

Durante le attività di riordino e inventariazione è emersa la scarsità della documentazione precedente al 1800 e afferente alla vita del Capitolo. Tuttavia, grazie ad alcuni antichi inventari sopravvissuti nell'Archivio della Chiesa di S. Lorenzo, è possibile conoscere parte dei fondi perduti e appartenuti all'Archivio capitolare laertino. È probabile che molti di questi documenti siano andati distrutti perché ritenuti ormai inutili o di nessuna valenza storica; altri, invece, sono andati dispersi o sottratti per varie cause o entrati a far parte di altri archivi pubblici e privati.
Fortunatamente, una parte consistente della documentazione smarrita è stata rintracciata presso l'Archivio privato della Famiglia dell'Aquila, dichiarato di notevole interesse storico con declaratoria del 25 gennaio 1978 [10]: si tratta di 5 inventari o "Stalloni di beni stabili e capitali" [11: 22-23] e diversi atti cartacei, rilegati con altri documenti dell'archivio privato, nonché una pergamena del 1572.
Il prezioso rinvenimento della documentazione in origine appartenuta al Capitolo laertino e le attività di riordino e inventariazione dell'Archivio privato dell'Aquila hanno permesso non solo di individuare le carte laertine prima considerate scomparse ma anche di indagare le motivazioni che provocarono la dispersione e lo smembramento dell'antico Archivio del Capitolo Collegiale della Chiesa Matrice di S. Lorenzo Martire di Laterza. In particolar modo, la ragione principale delle vicissitudini che caratterizzarono la vita dell'Archivio capitolare laertino risiede in una vertenza ultracentenaria sugli *Usi Civici* (1819-1955) sollevata dal Comune di Laterza per lo scioglimento delle promiscuità sul preteso demanio ecclesiastico laertino [6: 154-757, 761-767] e, nello specifico, la presenza cospicua della documentazione del Capitolo nell'Archivio privato dell'Aquila, a partire dal 1872-1875, trova spiegazione nell'acquisizioni da parte di tre fratelli dell'Aquila di beni di origine ecclesiastica, che erano oggetto da tempo della vertenza sugli *Usi Civici,* ereditandone la causa. Infatti, entrambi gli insiemi documentali, oggi conservati presso l'Archivio della Chiesa di S. Lorenzo e l'Archivio privato dell'Aquila, includono, tra gli altri, atti di acquisizioni, donazioni, legati e permute di beni patrimoniali oggetto del contenzioso.

3. METODOLOGIA E SFIDE

Stante lo stato dell'arte della documentazione fisica, primigenio obiettivo della ricerca è stato sperimentare la rigenerazione in ambiente digitale dell'originaria consistenza e organizzazione della serie, nell'ottica di ricomporre la documentazione e facilitare la fruizione e consultazione dei documenti dal punto di vista tipicamente archivistico.
Allo scopo, si è provveduto preliminarmente a digitizzare i documenti originali secondo gli standard e le buone pratiche in uso [3, 2], rispettando nella scansione la sequenza dei documenti allo stato dell'arte fisico. Quindi, i documenti delle due parti sono stati analizzati nelle loro componenti archivistiche per ricomporre l'organizzazione originaria della serie. Da questa analisi sono stati prodotti due inventari, nei quali si è data evidenza delle relazioni intercorrenti tra i documenti.
Fondamentali per rigenerare in ambiente digitale l'originaria consistenza della serie documentale sono risultate tre fonti primarie: l'*Inventario delle scritture e libri* redatto nel 1730 dal Clero e ancora presente tra le carte dell'Archivio di S. Lorenzo; l'*Elengo di testamenti e donazioni a favore del Capitolo* risalente alla metà dell'Ottocento, conservato presso l'Archivio privato dell'Aquila ma originariamente parte della documentazione del Capitolo, nel quale sono riportate brevi informazioni sui documenti prodotti dopo il 1730, inclusi alcuni oggi dispersi; lo *Stato dei Fondi Capitolari di Laterza*, un prospetto tabellare nel quale sono state riportare le registrazioni di atti notarili relativi ai beni posseduti dal Capitolo. Il riferimento a questi testimoni ha facilitato la ricostruzione digitale della serie per il periodo compreso tra la seconda metà del XVII e la metà del XIX secolo.
Durante l'analisi dei documenti, sono state rilevate diverse sequenze di fascicolazione e cartulazione, delle quali si è inteso identificare l'origine, le motivazioni che le hanno determinate e la funzione ai fini del riuso dei documenti. Nello specifico, sono state identificate 2 fascicolazioni e 6 cartulazioni apposte tra la fine del XVII e gli anni Trenta del XX secolo. Lo studio ha evidenziato che ciascuna sequenza non è dipesa da fattori legati alla gestione o riorganizzazione archivistica delle

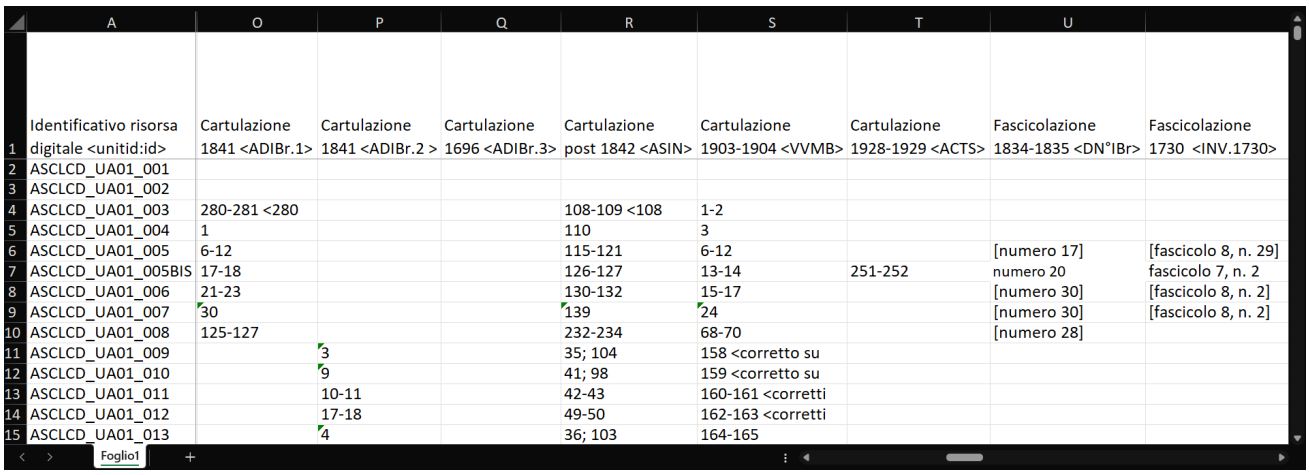
serie documentali, ma ha corrisposto alla necessità di creare aggregazioni dei documenti funzionali a differenti riusi in fasi specifiche della plurisecolare vertenza giudiziaria sugli *Usi Civici*.
Questo nuovo scenario ha indotto a studiare la possibilità di sperimentare ulteriori soluzioni digitali per consentire agli utenti di accedere e interagire con la documentazione secondo approcci non più esclusivamente archivistici, ma rispondenti a bisogni e ricerche peculiari agli studi storici, quali, a esempio, indagare le vicende storico-giuridiche che hanno caratterizzato il territorio laertino nell'arco di circa tre secoli, o esplorare i riusi che hanno determinato le diverse aggregazioni nelle varie fasi del contenzioso, spesso anche molto distanti tra loro nel tempo. La necessità di proporre organicamente in ambiente digitale una documentazione di tale complessità rendendola pienamente fruibile agli utenti ha reso necessario elaborare innanzitutto criteri di clusterizzazione dei vari sottoinsiemi documentali.
Nell'ottica di identificare ciascuna clusterizzazione per la gestione con i metadati, sono state utilizzate acronimizzazioni basate sulle diverse evidenze grafiche rilevate sulle carte, quali la posizione e la grafia delle numerazioni e il colore degli inchiostri, nelle more di definire criteri identificativi più stabili (vd. Tab. 1).

FASCICOLAZIONI		
ACRONIMO	DESCRIZIONE	DATAZIONE
INV.1730	Inventario 1730	1730
DN°IBr	Dorso, Numero, Inchiostro, Bruno	1834-1835

CARTULAZIONI		
ACRONIMO	DESCRIZIONE	DATAZIONE
ADIBr.1	Alto, Destra, Inchiostro, Bruno	1841
ADIBr.2	Alto, Destra, Inchiostro, Bruno	1841
ADIBr.3	Alto, Destra, Inchiostro, Bruno	1696
ASIN	Alto, Sinistra, Inchiostro, Nero	post 1842
VVMB	Vario, Vario, Matita Blu	1903-1904
ACTS	Alto, Centro, Talloncino a stampa	1928-1929

Tabella 1. Acronimizzazioni provvisorie delle clusterizzazioni documentali

La ricostruzione della prima fascicolazione settecentesca è stata clusterizzata con acronimo INV.1730. La seconda fascicolazione è stata identificata utilizzando per riscontro l'*Elengo* redatto dopo il 1834, oggi conservato tra le carte dell'Aquila (DN°IBr), che ha permesso di identificare i documenti che componevano l'aggregazione creata dal Clero tra il 1834 e il 1835 in seguito all'ordinanza emanata nel 1834 dal Consiglio d'Intendenza. La cartulazione ADIBr.3 identifica un sottoinsieme che include atti utilizzati in occasione di un contenzioso verificatosi nel 1696. Le cartulazioni ADIBr.1 e ADIBr.2 includono i documenti originariamente aggregati in quattro volumi di atti, presentati dal Clero all'Intendente di Terra d'Otranto nel 1841, mentre la cartulazione ASIN pertiene alla riproposizione degli stessi documenti insieme ad altri in una nuova aggregazione prodotta dopo il 1842. Le ultime due cartulazioni in ordine di tempo sono state apportate nel XX secolo da membri della famiglia dell'Aquila in occasione di riusi della documentazione antica legati a fasi tarde della vertenza demaniale, verificatesi la prima tra il 1903 e il 1904 (VVMB), la seconda tra il 1928 e il 1929 (ACTS).
L'obiettivo di rigenerare digitalmente la "vivacità", varietà e complessità della documentazione considerata ha guidato le scelte sui metadati da integrare nel tracciato in formato open data [1] da riutilizzare. La flessibilità e dinamicità del tracciato, infatti, rappresentano requisiti indispensabili sia per un'agile gestione della qualità e quantità di elementi e attributi da includere, sia per customizzare le risorse digitali in relazione alle tipologie di oggetti da gestire e agli obiettivi di fruizione e interazione in prospettiva *user-oriented*. Infine, metadatare in formato open, se eseguito in conformità con le regole e le buone pratiche di riferimento, garantisce la piena coerenza con i principi FAIR [13], favorendo al massimo potenziale la ricercabilità, accessibilità, interoperabilità e riusabilità delle risorse digitali.
Il dataset, infatti, è stato progettato e realizzato per essere riversato nella DL *Open Memory*. A tal fine, il tracciato precedentemente utilizzato è stato rielaborato includendo metadati atti ad attivare modalità accesso alle risorse a vari livelli: per consultare la serie nella sua originaria organizzazione nell'Archivio Capitolare, oppure per accedere separatamente ai due insiemi identificati con le sedi di conservazione attuali, o, ancora, per consultare direttamente le singole aggregazioni in base ai riusi accedendo direttamente alle varie cartulazioni. Nello specifico, l'inserimento di metadati di relazione di contesto per gestire le singole clusterizzazioni (vd. Fig. 2) è risultato fondamentale nella prospettiva dello sviluppo dell'interfaccia di interrogazione, in quanto garantiscono l'accesso alle informazioni sulle diverse fascicolazioni e cartulazioni, rappresentando le chiavi necessarie a esplorare il complesso documentale dal punto di vista storico-giuridico e culturale, cosa impossibile sulla documentazione fisica.



	A	O	P	Q	R	S	T	U	
1	Identificativo risorsa digitale <unitid:id>	Cartulazione 1841 <ADIBr.1>	Cartulazione 1841 <ADIBr.2 >	Cartulazione 1696 <ADIBr.3>	Cartulazione post 1842 <ASIN>	Cartulazione 1903-1904 <VVMB>	Cartulazione 1928-1929 <ACTS>	Fascicolazione 1834-1835 <DN°IBr>	Fascicolazione 1730 <INV.1730>
2	ASCLCD_UA01_001								
3	ASCLCD_UA01_002								
4	ASCLCD_UA01_003	280-281 <280			108-109 <108	1-2			
5	ASCLCD_UA01_004	1			110	3			
6	ASCLCD_UA01_005	6-12			115-121	6-12		[numero 17]	[fascicolo 8, n. 29]
7	ASCLCD_UA01_005BIS	17-18			126-127	13-14	251-252	numero 20	fascicolo 7, n. 2
8	ASCLCD_UA01_006	21-23			130-132	15-17		[numero 30]	[fascicolo 8, n. 2]
9	ASCLCD_UA01_007	30			139	24		[numero 30]	[fascicolo 8, n. 2]
10	ASCLCD_UA01_008	125-127			232-234	68-70		[numero 28]	
11	ASCLCD_UA01_009		3		35; 104	158 <corretto su			
12	ASCLCD_UA01_010		9		41; 98	159 <corretto su			
13	ASCLCD_UA01_011		10-11		42-43	160-161 <corretti			
14	ASCLCD_UA01_012		17-18		49-50	162-163 <corretti			
15	ASCLCD_UA01_013		4		36; 103	164-165			

Foglio1

Figura 2. Metadati per la gestione delle clusterizzazioni documentali

In tal modo, interagendo con la collezione digitale mediante interfaccia di interrogazione, l'utente può accedere a una singola cartulazione indicizzata e ottenere in risposta i documenti appositamente aggregati per affrontare un momento specifico della vertenza demaniale, nonché porli in relazione con gli altri insiemi in cui gli stessi documenti sono stati uniti a ulteriori evidenze in base alle differenti necessità, consultando una o più fasi del contenzioso giudiziario.

4. CONCLUSIONI

La ricerca che qui si presenta apre nuovi scenari nella fruizione di beni documentali in ambiente digitale.
La sperimentazione avviata sui due complessi documentali ha reso possibile riorganizzare in digitale le carte d'archivio conservate in sedi fisiche differenti, ricomponendo una serie archivistica andata smembrata nel corso dei secoli per diverse vicende storiche. I risultati della ricerca, in realtà, hanno aperto un ulteriore scenario di interazione con la documentazione, che possa produrre nuovi interessi storico-culturali. Infatti, elementi come le cartulazioni e altre peculiarità della documentazione archivistica, solitamente trascurati nelle analisi di fonti storiche, hanno mostrato una rinnovata visione della ricerca, in quanto diventano testimonianze di usi e riusi che i documenti hanno vissuto nel corso del tempo, non in relazione a ordinamenti strettamente archivistici ma riconducibili a motivazioni pratiche legate alle vicissitudini che hanno caratterizzato l'intero ciclo di vita dei documenti. Il digitale consente, dunque, di fruire della raccolta documentale attraverso le funzioni pratiche svolte dai documenti, scoprendo storie che raccontano circostanze altrimenti non conosciute. Quest'ultimo aspetto ha generato nuove prospettive nella sperimentazione di interfacce di interazione utente innovative, oggetto di ricerca attualmente in corso, poiché metadatazioni accurate possono favorire l'accessibilità alle risorse e alle collezioni digitali da diversi punti di vista.

BIBLIOGRAFIA

[1] Aliprandi, Simone. *Il Fenomeno Open Data. Indicazioni e norme per un mondo di dati aperti*. Milano: Ledizioni, 2014.

[2] Allegrezza, Stefano. *La digitalizzazione del patrimonio culturale. Linee guida, standard, esperienze*. Torre del Lago, Lucca: Civita Editoriale, 2021.

[3] Barbuti, Nicola. *La digitalizzazione dei beni documentali metodi, tecniche, buone prassi.* Milano: Editrice Bibliografica, 2022.

[4] Barbuti, Nicola. «Ripensare i formati, ripensare i metadati: prove "tecniche" di conservazione digitale». *Umanistica Digitale* 5 (2019). https://doi.org/10.6092/issn.2532-8816/9055.

[5] Barbuti, Nicola, Giuliano De Felice, Annalisa Di Zanni, Paolo Russo, e Altheo Valentini. «Creating Digital Culture by Co-Creation of Digital Cultural Heritage: The Crowddreaming Living Lab Method». *Umanistica Digitale* 9 (2020): 19–34. https://doi.org/10.6092/issn.2532-8816/9956.

[6] Bollettino dei demani comunali delle provincie meridionali continentali: sentenze della Commissione feudale, ordinanze, decreti ed altri atti della sistemazione dei demani. Vol. 2. Roma: Tipografia Nazionale di G. Bertero e C., 1911.

[7] Bongermino, Raffaella. *Storia di Laterza: gli eventi, l'arte, la natura*. Galatina: Congedo, 1993.

[8] De Bari, Mauro. «Verso una fruizione phygital del Patrimonio Culturale in termini di OpenCulture». In *Il patrimonio culturale pugliese. Ricerche, applicazioni e best practices. Atti del II congresso Beni culturali in Puglia. Bari 28-30 settembre 2022*, a cura di Giovanna Fioretti e Cinzia Campobasso, 182–86. Fondazione Pasquale Battista, 2023.

[9] De Bari, Mauro, e Nicola Barbuti. «Addressing User Engagement With an Interactive Reading Model by Innovative Digital Expansion». In *Proceedings of the 18th Italian Research Conference on Digital Libraries. Padua 24-25 February 2022*, a cura di

Giorgio Maria Di Nunzio, Beatrice Portelli, Domenico Redavid, e Gianmaria Silvello, 3160:1–8. CEUR Workshop Proceedings, 2022.

[10] Dell'Aquila, Carlo. *Bibliografia laertina*. Bari: Centro di ricerche storiche della Pro Loco di Laterza, 1978.

[11] Dell'Aquila, Carlo. *Laterza Sacra*. Taranto: Amministrazione Provinciale (Manduria, Tiemme), 1989.

[12] Dell'Aquila, Carlo. (a cura di). *Per la storia di Laterza. Fonti archivistiche e documentarie*. Galatina: Congedo, 1993.

[13] Hermon, Sorin, e Franco Niccolucci. «FAIR Data and Cultural Heritage Special Issue Editorial Note». *International Journal on Digital Libraries* 22 (2021): 251–55. https://doi.org/10.1007/s00799-021-00309-8.

[14] Tatò, Grazia. «Archivio della chiesa di San Lorenzo Martire in Laterza. Inventario». In *Per la storia di Laterza. Fonti archivistiche e documentarie*, a cura di Carlo Dell'Aquila, 17–56. Galatina: Congedo, 1993.

Preservare e valorizzare la memoria di archivi storici di ex-ospedali psichiatrici

Grazia Serratore
Università della Calabria e IIT-CNR, Rende (CS), Italia - grazia.serratore@iit.cnr.it

ABSTRACT

Questo contributo presenta attività di ricerca inerenti alle cartelle cliniche degli ex ospedali psichiatrici calabresi di Reggio Calabria e Girifalco (CZ). L'obiettivo è illustrare lo stato di conservazione dei due archivi storici e proporre nuove prospettive di ricerca, supportate dall'impiego di metodologie di intelligenza artificiale e da tecniche di diagnostica non invasive, per favorire l'accesso al patrimonio informativo e la conservazione nel tempo delle carte. Tali attività, che si inquadrano nell'ambito del progetto di dottorato descritto, si propongono di completare le fasi di i) digitalizzazione e trascrizione delle cartelle cliniche dei due archivi storici, avviate nel corso di precedenti lavori e progetti di ricerca; ii) realizzare strumenti per favorire l'organizzazione, l'accesso e la consultazione del materiale archivistico; iii) effettuare valutazioni sullo stato di conservazione delle cartelle cliniche.

Le scansioni delle cartelle cliniche non consentono un accesso immediato ed interoperabile ai contenuti delle stesse, per cui si applicheranno tecniche di Handwritten Text Recognition, mediante l'uso del software Transkribus, per ottenere delle trascrizioni in formato machine-readable. Il software servirà per addestrare un modello specifico di riconoscimento del testo manoscritto delle cartelle cliniche, redatte in lingua italiana e relative al XIX e al XX secolo. Saranno presi in considerazione solo le tabelle nosologiche, i diari delle degenze e i testi elaborati dai pazienti, sia per il particolare interesse che queste tipologie documentali rivestono per fini di ricerca e sia perché l'assenza di dati anagrafici al loro interno consente un maggiore rispetto della riservatezza e della privacy. Le difficoltà principali nell'addestramento del modello risiedono nella presenza di termini specialistici, in alcuni casi anche desueti, della disciplina psichiatrica, di diverse calligrafie all'interno di una stessa pagina e nell'evoluzione della struttura della cartella clinica, soprattutto per quanto riguarda lo sviluppo delle sezioni della tabella nosologica. Le trascrizioni rappresenteranno il punto di partenza per le indagini successive: si applicheranno tecniche di Natural Language Processing per automatizzare i processi di indicizzazione e categorizzazione dei contenuti e la Network Analysis per snellire le fasi di costruzione di un Knowledge Organization System (KOS).

PAROLE CHIAVE

Archivi storici; ex ospedali psichiatrici; Natural Language Processing; indicizzazione automatica; Knowledge Organization Systems.

1. INTRODUZIONE

Gli archivi degli ex ospedali psichiatrici costituiscono una preziosa fonte di conoscenza perché in essi «è scritta più che altrove la storia della sofferenza dei singoli e delle comunità, la storia della solidarietà, del controllo sociale e delle cure» [27: 43]. Negli anni successivi alla chiusura di queste istituzioni, in seguito all'entrata in vigore della Legge Basaglia[1], è stata posta molta attenzione al recupero della memoria contenuta negli archivi storici di ex ospedali psichiatrici [2, 12]. Fra le strutture manicomiali dell'Italia meridionale ha rivestito particolare importanza l'ospedale psichiatrico di Girifalco (CZ), sia per la sua collocazione geograficamente strategica, che ha permesso di alleviare l'affollamento della Real Casa De' Matti di Aversa (CE), all'epoca il manicomio più a sud della penisola; sia per i diversi e innovativi approcci di cura voluti dai vari luminari che si sono succeduti alla guida dell'istituto. La struttura ha accolto i primi pazienti nel 1881 e, in quasi un secolo di attività, ha contato ben 22.415 ingressi [20]. Le cartelle cliniche ad oggi presenti in archivio sono circa 15.800. Questa discrepanza numerica è dovuta ad accessi multipli dello stesso paziente o allo smarrimento e al deterioramento di alcune carte.

Nel 1932 è stata aperta la seconda struttura manicomiale calabrese a Reggio Calabria, che annovera tra i suoi degenti anche il brigante Musolino. Ad oggi l'archivio storico contiene circa 12.500 cartelle cliniche.

Le cartelle cliniche, soprattutto quelle più datate, sono riccamente compilate e risentono degli eventi storici e dei fattori sociali contingenti alla loro stesura. Queste rappresentano un contenitore disomogeneo di informazioni, in cui è possibile

[1] Entrata in vigore il 13 maggio 1978, la legge n. 180/78, passata alla storia come legge Basaglia, sancì la definitiva chiusura dei manicomi, riformando il sistema di cura per il disagio mentale e segnando una svolta nel mondo dell'assistenza ai pazienti psichiatrici, si veda: https://www.gazzettaufficiale.it/eli/id/1978/05/16/078U0180/sg

trovare non solo notizie relative allo stato di salute del paziente o dati anagrafici e amministrativi, ma anche la corrispondenza intercorsa tra il paziente e i propri familiari nel periodo di internamento o alcuni frammenti manoscritti composti dal paziente. In virtù di questa varietà e ricchezza di contenuti, «la cartella clinica consente di adottare una pluralità di approcci nello studio delle pratiche di internamento perché al suo interno è possibile rintracciare scorci di storia e storie che derivano dall'incontro tra il sapere psichiatrico e i codici culturali della popolazione internata» [15].
Inoltre, le carte sono soggette a un naturale e continuo deterioramento, legato ai processi di ossidazione e idrolisi, provocati da fattori ambientali e da agenti atmosferici, e all'azione di agenti infestanti (batteri, funghi, ecc.). Per preservarle è necessario uno studio preliminare sulle caratteristiche dei materiali che costituiscono le carte stesse, sull'ambiente in cui si trovano e sullo stato di conservazione [8], in modo da intervenire con attività di diagnostica e individuare eventuali fattori di rischio e, di conseguenza, valutare la necessità di intervento e le tecniche di conservazione preventiva più idonee da adottare. Questa tipologia di analisi consente di definire un protocollo di conservazione dei supporti cartacei e di generare una conoscenza integrata tra le competenze di dominio archivistico e le metodologie di conservazione preventiva. Dunque, se opportunamente valorizzati, gli archivi degli ex ospedali psichiatrici di Girifalco e di Reggio Calabria possono rappresentare un osservatorio unico per analizzare da diverse prospettive disciplinari l'esperienza manicomiale calabrese nel XIX e nel XX secolo. Questo contributo intende, quindi, illustrare lo stato di conservazione dei due archivi storici e proporre nuove prospettive di ricerca, supportate dall'impiego di metodologie di intelligenza artificiale e da tecniche di diagnostica non invasive, per favorire l'accesso al patrimonio informativo e la conservazione nel tempo delle carte[2].

2. STATO DELL'ARTE

Gli archivi dei due ex ospedali psichiatrici sono stati oggetto di vari studi e progetti. In particolare, occorre menzionare il progetto Carte da Legare[3], condotto dalla Direzione Generale Archivi (DGA) che, nel tentativo di raccordare singole iniziative, si propone di recuperare e salvaguardare il patrimonio archivistico degli ex ospedali psichiatrici italiani, articolandosi su tre livelli di intervento: il censimento degli archivi, il riordinamento e l'inventariazione della documentazione. Grazie al lavoro condotto fino ad ora, è possibile consultare online gli inventari di vari istituti e condurre delle valutazioni statistiche sulla base dei dati relativi alle cartelle cliniche schedate per ogni istituto. L'archivio del manicomio di Girifalco è parzialmente inventariato in Carte da Legare, grazie al lavoro di riconciliazione e riversamento dei metadati delle cartelle cliniche condotto dalla sede di Cosenza dell'Istituto di Informatica e Telematica del Consiglio Nazionale delle Ricerche (IIT-CNR). L'archivio del manicomio di Reggio Calabria, invece, non è censito, in quanto è attualmente in fase di inventariazione.
Le cartelle cliniche dell'ex ospedale psichiatrico di Girifalco sono state oggetto di importanti lavori di ricerca scientifica [10, 17, 18], tra cui assume particolare rilievo lo studio genealogico condotto dal Centro Regionale di Neurogenetica (CRN) di Lamezia Terme (CZ) per rilevare l'ereditarietà di alcune malattie neurodegenerative, come la malattia di Alzheimer, e che ha permesso di retrodatare la prima attestazione clinica di questa patologia [3, 9]. Il progetto ALPHA (eAsy inteLligent service Platform for Healthy Ageing)[4], svolto in collaborazione dall'IIT e dall'ICAR (Istituto di Calcolo e Reti ad Alte Prestazioni) del CNR e dal CRN, ha avuto l'obiettivo di definire profili di rischio dell'insorgenza di patologie neurodegenerative, mediante un'analisi dei segni e dei sintomi presenti nelle cartelle cliniche dei pazienti con uno stato di Mild Cognitive Impairment, conducendo al contempo un'analisi diacronica sui segni e sintomi dei pazienti affetti da patologie dementigene ricoverati a Girifalco. Nello svolgimento del progetto sono state digitalizzate le cartelle cliniche relative al periodo 1881-1931 e le prime 500, relative al periodo 1881-1894, sono state anche trascritte.
In letteratura è possibile trovare diversi contributi volti ad analizzare l'evoluzione storica della disciplina psichiatrica e l'esperienza della detenzione manicomiale in Italia nell'Ottocento e nel Novecento [1, 13], facendo spesso riferimento a delle singole realtà locali e ponendo l'attenzione su casi particolari, come la condizione femminile e dei minori nei manicomi [14, 26]. In particolare, nel solco di quest'ultimo filone di ricerche, è stato condotto uno studio sui minori ospitati nel manicomio di Girifalco, analizzando le trascrizioni già disponibili delle 500 cartelle cliniche relative al periodo 1881-1894 [5]. Inoltre, sullo stesso campione sono stati condotti studi linguistici finalizzati all'individuazione e all'analisi del lessico della psichiatria, in particolare sulla terminologia utilizzata per indicare gli strumenti e le terapie adoperati nel corso

[2] Le cartelle cliniche di ex ospedali psichiatrici contengono al loro interno dati sensibili, legati non solo allo stato di salute dei pazienti, ma in alcuni casi anche alle loro convinzioni religiose e filosofiche, per questa ragione, nella loro consultazione e nelle attività di ricerca è importante assicurare la tutela della riservatezza e della privacy in ottemperanza a quanto prescritto dal Regolamento Generale sulla Protezione dei Dati (GDPR) in materia di protezione dei dati personali.
[3] Per conoscere maggiori dettagli sul progetto "Carte da legare. Archivi della psichiatria in Italia" si veda: https://cartedalegare.cultura.gov.it/home
[4] Maggiori informazioni sul progetto ALPHA (eAsy inteLligent service Platform for Healthy Ageing) sono disponibili al: http://alpha.iit.cnr.it/alpha/

della pratica clinica e sulla terminologia calda e fredda della psichiatria, in un periodo in cui questa veniva utilizzata senza un valore mono-referenziale [6, 7, 10, 11].
Gli archivi storici degli ex ospedali psichiatrici offrono dunque diverse prospettive di ricerca, non ultima quella relativa all'organizzazione e alla rappresentazione della conoscenza specialistica: il dominio della psichiatria è caratterizzato dalla presenza di differenti sistemi di organizzazione della conoscenza che, con livelli di complessità diversi, si propongono di strutturare i concetti del settore. Questi nascono principalmente in ambito medico per agevolare e per standardizzare la comunicazione in materia di salute mentale e, in particolare, è possibile annoverare: il sistema nosografico Diagnostic and Statistical Manual of Mental Disorders (DSM-5); il sistema di classificazione International Classification of Diseases (ICD); la Hierarchical Taxonomy Of Psychopathology (HiTOP); il vocabolario controllato Medical Subject Headings (MeSH); il glossario multilingue "Diccionari de psiquiatria".
Da una prima ricognizione effettuata non risultano sistemi di organizzazione della conoscenza appartenenti esclusivamente al dominio psichiatrico che siano nativi in lingua italiana, per cui è possibile prendere in considerazione solo risorse che trattano il dominio della psichiatria parzialmente e non come oggetto principale, quali il Thesaurus del Nuovo Soggettario[5], usato per l'indicizzazione per soggetto delle risorse in ambito bibliotecario, e il Tesauro italiano di bioetica, strumento di accesso alla letteratura bioetica e biomedica, attraverso l'individuazione e la descrizione delle parole-chiave dei testi sull'argomento. Per tutti gli strumenti citati sarà necessario effettuare una valutazione più precisa per comprendere la loro applicabilità al contesto di studio, in merito ad esempio alla copertura semantica, alla specificità dei termini e alla struttura classificatoria adottata.

3. FASI DI LAVORO E METODOLOGIA

Le cartelle cliniche degli ex ospedali psichiatrici, in quanto beni culturali, sono soggette a tutela da parte della DGA e devono essere preservate da fenomeni di degrado della carta. Il processo di digitalizzazione, oltre ad agevolare la loro fruizione e a garantire la presenza di un corrispettivo digitale, assicura che queste siano tutelate e conservate nel lungo periodo[6]. Infatti, un archivio digitale, costituito dalle copie per immagine delle cartelle cliniche, consente di accedere ai contenuti documentali di interesse senza dover consultare direttamente gli originali cartacei, preservandoli dai fenomeni di degrado legati al ripetuto maneggiamento. Tuttavia, le scansioni delle cartelle cliniche, da sole, non consentono un accesso immediato al contenuto documentale. Per questa ragione, è necessario convertire le immagini in testo machine-readable, applicando ad esempio tecniche di Handwritten Text Recognition. I lavori precedentemente menzionati hanno permesso di avere a disposizione parte del complesso archivistico di Girifalco già digitalizzato (5.258 cartelle cliniche) e parte del digitalizzato già trascritto (550 cartelle cliniche) mediante un software di dettatura vocale. L'archivio storico del manicomio di Reggio Calabria è, invece, solo parzialmente digitalizzato (219 registri).
Il progetto di dottorato descritto nel presente lavoro si propone, pertanto, di completare le fasi di digitalizzazione e trascrizione, di realizzare strumenti per favorire l'organizzazione, l'accesso e la consultazione del materiale e di realizzare alcune ricerche di carattere storico-linguistico.
Le cartelle cliniche sono in realtà dei fascicoli personali dei pazienti, che contengono al loro interno anche atti disposti e raccolti non per finalità di cura, ma per scopi amministrativi. In particolare, è possibile trovare: il certificato di nascita; la modula informativa, ovvero un documento prodotto dal medico del comune di provenienza per attestare le condizioni antecedenti il ricovero e eventuali altre notizie riguardanti la storia clinica del paziente e dei suoi parenti prossimi; l'ordine di ricovero emanato dall'autorità politica competente; l'atto notorio con attestazione della diagnosi; alcuni documenti comprovanti lo status economico di povertà assoluta o relativa del ricoverato e dei suoi parenti; l'elenco degli oggetti preziosi in possesso del paziente al momento dell'ammissione; la tabella nosologica; il diario clinico e altre annotazioni di carattere sanitario; lettere e biglietti scritti dai pazienti.
Tutta la documentazione presente all'interno di ogni cartella clinica verrà digitalizzata, in modo da poter disporre in formato elettronico dell'intero archivio, e il processo di digitalizzazione sarà effettuato nei locali in cui l'archivio è preservato, mediante scanner portatili a testina mobile.

[5] Sono stati condotti alcuni studi sull'impiego del Nuovo soggettario nella descrizione semantica di fondi archivistici, conducendo così una mappatura tra il Nuovo Soggettario e alcuni Thesauri specialistici e prevedendo l'acquisizione della terminologia mancante nel Thesaurus del Nuovo soggettario. Ne sono un esempio le attività condotte dal Gruppo Linguaggi del MAB Toscana per rendere possibile l'indicizzazione per soggetto e la descrizione delle risorse documentarie non bibliografiche (archivistiche e museali) con il Nuovo Soggettario, come nel caso dell'Archivio Ernesto Rossi e del Museo Galileo. Tra i progetti più recenti si annovera PerformArt, che ambisce ad arricchire la comprensione della storia delle arti performative presso la nobiltà romana tra il 1644 e il 1740 utilizzando la documentazione contenuta negli archivi di undici famiglie aristocratiche.
[6] Quanto detto è in linea con gli obiettivi prefissati dal Piano Nazionale di Digitalizzazione redatto dall'Istituto Centrale per la Digitalizzazione del Patrimonio Culturale – *Digital Library* del Ministero della Cultura.

Il processo di trascrizione del testo manoscritto sarà effettuato sfruttando il potenziale offerto dall'intelligenza artificiale e, nello specifico, verrà utilizzato il software Transkribus, un tool per il riconoscimento automatico della scrittura. Si effettuerà la trascrizione solo delle tabelle nosologiche, dei diari clinici e della produzione manoscritta dei pazienti. Questa scelta è legata non solo al valore che queste tipologie documentali rivestono per fini di ricerca, poiché contengono dettagli sulla sintomatologia, sulle caratteristiche psico-fisiche dei pazienti e sulle terapie praticate, ma soprattutto per lo sporadico utilizzo al loro interno di nomi propri, dati anagrafici e dati personali.
Al contrario il frontespizio, l'ordine di ricovero, l'atto notorio e la dichiarazione dello status economico contengono campi valorizzati con informazioni sensibili e dati particolari che, in molti casi, consentono di identificare univocamente un degente e alcuni membri della sua famiglia o suoi conoscenti. Sebbene le politiche sull'utilizzo di Transkribus ribadiscano che le immagini e i dati utilizzati sulla piattaforma sono sempre conservati sui server di READ-COOP SCE in modo conforme a quanto stabilito dal Regolamento generale per la protezione dei dati personali (Regolamento UE 2016/679), vista la natura dei dati trattati, si è volutamente scelto di elaborare solo le pagine i cui contenuti non consentano di identificare in nessun caso un degente.
Transkribus fornisce dei modelli generali per il riconoscimento del testo manoscritto in diverse lingue e l'unico disponibile per la lingua italiana è Transkribus Italian Handwriting M1, che fa riferimento al periodo dal XVI al XVII secolo. Al fine di testare il modello sulla documentazione oggetto di analisi sono state trascritte cinque pagine ma, a causa della diversa natura temporale e della specifica tipologia documentale, i risultati ottenuti hanno dimostrato la necessità di addestrare un modello specifico. A tal fine, il tool necessita di una fase di apprendimento supervisionato, per cui è necessario un dataset etichettato, chiamato ground-truth, da cui possa apprendere come interpretare ogni riga dell'immagine e riportarla in una riga di testo in output [21]. Nel caso specifico, disponendo già di un buon numero di trascrizioni, si è deciso di associare il testo già disponibile alle rispettive immagini delle cartelle cliniche in modo da costruire la ground-truth del modello. Associare ogni riga di testo alla corrispettiva riga dell'immagine è un'attività onerosa, ma sicuramente non quanto elaborare una trascrizione *ex novo*.
Il software Transkribus consentirà non solo di compiere una funzione di Handwritten Text Recognition, ma anche di addestrare nuovi modelli per riconoscere automaticamente dei layout complessi e condurre una Layout Analysis, che sarà utile per valutare l'evoluzione strutturale delle cartelle cliniche nel tempo e per avvalorare il riscontro di eventuali mutamenti emersi dallo spoglio manuale effettuato su campioni di cartelle cliniche.
Il layout del diario clinico rimane pressoché stabile: si passa da una struttura tabellare, in cui sono riportate le date delle annotazioni, le osservazioni fatte dai sanitari e alcune prescrizioni di carattere igienico, con intestazione "Diario Clinico" o "Decorso della cura", ad una pagina a righe vuota con la sola intestazione "Diario Clinico (durante la degenza nell'Ospedale)". La tabella nosologica, invece, ha subito delle variazioni strutturali più evidenti: fino ai ricoveri avvenuti nel 1883 presenta l'intestazione "Manicomio provinciale della Calabria ulteriore seconda in Girifalco" e prevede il campo 'Diagnosi frenopatica", mentre successivamente riporta l'intestazione "Manicomio provinciale di Catanzaro in Girifalco" con il campo "Diagnosi della pazzia" e, poi, soltanto "Diagnosi". In particolare, la tabella nosologica contiene la sezione "Stato presente", nelle cui sottosezioni vengono descritte le funzioni psichiche, le funzioni di relazione, le funzioni fisiche e le "misure antropologiche". Quest'ultima, con il passare degli anni, sarà sempre più articolata e assumerà la denominazione "Note antropologiche degenerative".
La tabella nosologica risulta generalmente compilata dallo stesso autore, in quanto redatta presumibilmente al momento del ricovero, mentre il diario clinico presenta diverse calligrafie, poiché in molti casi ripercorre la degenza di anni nell'ospedale psichiatrico.
La lingua utilizzata per la stesura delle cartelle cliniche è l'italiano del XIX e del XX secolo, ma a seconda del livello di istruzione di chi scrive l'aderenza alle norme grammaticali dell'epoca è variabile.
Si partirà dalle trascrizioni per applicare processi di indicizzazione ed estrazione automatica di relazioni semantiche. Per agevolare gli utenti finali nel reperire informazioni rilevanti, rispetto ai propri obiettivi di ricerca, è necessario strutturare la conoscenza implicita nei documenti indicizzando i contenuti delle cartelle cliniche. Il processo di indicizzazione può essere definito come l'azione di descrivere o identificare un documento con riferimento al suo contenuto concettuale (ISO 5963:1985) e, in ambito documentale, indica la selezione, in un linguaggio controllato, di descrittori formali utilizzabili come punti di accesso per concetti e non per termini ai documenti [25]. Nel caso di specie, considerando che il tradizionale processo di indicizzazione manuale è molto oneroso in termini di tempo e di risorse, dovendo lavorare con una grande mole di documenti, si opterà per un processo di indicizzazione automatica delle cartelle cliniche. I termini maggiormente rappresentativi verranno selezionati non solo sulla base di criteri statistici, legati ad esempio alla frequenza o alla funzione di peso TF-IDF, ma saranno scelti sulla base del contesto di occorrenza e della loro effettiva capacità di rendere una cartella clinica riconoscibile tra le altre. Inoltre, il contesto di occorrenza dei termini all'interno delle cartelle cliniche sarà utile per tenere traccia delle diverse sfumature di significato nel tempo. I descrittori formali selezionati dovranno essere

opportunamente gestiti, in modo da ricondurre tutti i sinonimi o le forme varianti di un termine ad un unico termine preferito così da migliorare l'Information Retrieval e gestire il rumore informativo, modulando la precisione e il richiamo [24]. Dopo le fasi di tokenizzazione e lemmatizzazione, i termini rappresentativi estrapolati dai documenti saranno ricondotti a dei termini preferiti selezionati secondo criteri di vicinanza concettuale alle forme e alle parole chiave che effettivamente potrebbero essere impiegati da un utente in fase di ricerca. Ad esempio, nelle cartelle cliniche è possibile trovare "giubbetto di sicurezza" per indicare il mezzo di contenimento utilizzato per assicurare i pazienti particolarmente agitati, ma un utente moderno effettuerebbe la sua ricerca verosimilmente a partire da "camicia di forza".
Nella fase di selezione, quindi, si terrà conto delle forme utilizzate nelle cartelle cliniche poiché, da un punto di vista linguistico, queste rappresentano importanti attestazioni storiche del nascente linguaggio della psichiatria e verranno mantenute come termini non preferiti. I termini non preferiti non saranno utilizzati per indicizzare i documenti, ma costituiranno dei punti di accesso da cui un utente può essere indirizzato al termine preferito.
Le motivazioni legate a questa scelta sono la necessità di gestire la presenza di termini diversi per indicare il medesimo concetto all'interno di cartelle cliniche diverse e in periodi differenti. In questo modo, sarà possibile valorizzare e rendere più accessibile il patrimonio documentale agli utenti finali, grazie alla costruzione di percorsi di ricerca intuitivi e vicini al loro uso reale della lingua.
Per automatizzare la fase di indicizzazione, verranno usate le librerie Python più idonee per applicare tecniche di Natural Language Processing [16, 19, 23], come ad esempio Named Entity Recognition e Network Analysis, e di Deep Learning.
La tecnica di Named Entity Recognition sarà utile per estrapolare direttamente dal testo luoghi, riferimenti temporali e altre entità nominali rappresentative [22], mentre la Network Analysis permetterà, ad esempio, di avere una visualizzazione delle relazioni complesse che legano gli eventi clinici, le varie patologie con le cure adottate all'epoca, consentendo di identificare pattern tra sintomi specifici, trattamenti prescritti, risultati dei trattamenti. Infatti, i termini estrapolati come particolarmente rappresentativi dei concetti andranno classificati e ricondotti a delle categorie ben precise, in modo da poter poi individuare le relazioni semantiche che li legano e che garantiranno la possibilità di generare connessioni reciproche tra le informazioni.
L'indicizzazione e l'estrazione delle relazioni semantiche verranno svolte automaticamente, con il fine di snellire le fasi di strutturazione della conoscenza contenuta nelle cartelle cliniche e, soprattutto, con l'intento di costruire un Knowledge Organization System (KOS) che garantisca il recupero delle informazioni di interesse e consenta di rendere più accessibili e fruibili i due archivi storici presi in considerazione.
A causa dell'eterogeneità dei KOS disponibili, sarà necessario effettuare una valutazione iniziale delle sovrapposizioni e delle differenze che questi presentano sia nelle finalità di progettazione sia nella copertura semantica. La lacuna riscontrata per i sistemi di organizzazione della conoscenza in lingua italiana applicabili al dominio della psichiatria potrebbe rappresentare una prospettiva di ricerca futura per disporre, anche all'interno del contesto nazionale italiano, di un sistema finalizzato all'organizzazione della conoscenza nel settore psichiatrico, che sia non solo un punto di riferimento per gli esperti del settore, ma anche un ausilio nel processo di traduzione degli altri KOS di dominio già disponibili in altre lingue.
In parallelo verranno svolte anche le valutazioni sullo stato di conservazione delle cartelle cliniche, soprattutto per quelle di particolare interesse storico, applicando esclusivamente tecniche diagnostiche non distruttive e non invasive, come ad esempio la spettroscopia Raman [4]. Queste indagini preliminari serviranno per identificare gli elementi costitutivi della carta ed individuare dei campioni simili da sottoporre ad uno studio di invecchiamento monitorato. I risultati ottenuti consentiranno di definire delle strategie per prolungare la vita del patrimonio documentale.

4. CONCLUSIONI

La digitalizzazione delle cartelle cliniche garantisce la tutela e la valorizzazione di questo prezioso patrimonio documentale cartaceo. Nella costruzione di un archivio digitale occorre prestare particolare attenzione agli aspetti legati alla conservazione a norma nel lungo periodo, quali ad esempio la scelta dei formati più stabili e robusti e il costante aggiornamento delle tecnologie utilizzate per la loro gestione, così da evitare i problemi legati all'obsolescenza tecnologica.
Inoltre, processi quali l'indicizzazione e l'estrazione automatica di relazioni semantiche consentiranno di strutturare la conoscenza latente nelle cartelle cliniche, aprendo, come nel caso dei due archivi storici degli ex ospedali psichiatrici di Girifalco e Reggio Calabria, la strada ad innumerevoli ed inediti percorsi di ricerca che gli utenti finali potranno intraprendere senza la necessità di doversi recare fisicamente in archivio per ottenere le informazioni di interesse.
La tutela e la valorizzazione del patrimonio documentale devono, quindi, necessariamente essere finalizzate anche ad un miglioramento delle possibilità di fruizione, solo in questo modo è possibile costruire una conoscenza circolare.

BIBLIOGRAFIA

[1] Babini, Valeria Paola. *Liberi tutti: manicomi e psichiatri in Italia una storia del Novecento*. Bologna: Il Mulino, 2011.

[2] Baggio, Serenella. «Memorie di guerra degli archivi manicomiali del Trentino». In *Lingua e patologia. I sistemi instabili*, di Francesca Maria Dovetto, 203–34. 5. Roma: Aracne Editrice - Gioacchino Onorati Editore, 2020.

[3] Borrello, Laura, Chiara Cupidi, Valentina Laganà, Maria Anfossi, Maria Elena Conidi, Nicoletta Smirne, Maria Taverniti, Roberto Guarasci, e Amalia Cecilia Bruni. «Angela R.: A Familial Alzheimer's Disease Case in the Days of Auguste D.» *Journal of Neurology* 263, fasc. 12 (dicembre 2016): 2494–98. https://doi.org/10.1007/s00415-016-8294-x.

[4] Botti, Sabina, Francesca Bonfigli, Valentina Nigro, Alessandro Rufoloni, e Angelo Vannozzi. «Evaluating the Conservation State of Naturally Aged Paper with Raman and Luminescence Spectral Mapping: Toward a Non-Destructive Diagnostic Protocol». *Molecules* 27, fasc. 5 (5 marzo 2022): 1712. https://doi.org/10.3390/molecules27051712.

[5] Chiaravalloti, Maria Teresa, e Maria Taverniti. «Sanus egredieris». *Mélanges de l'École française de Rome - Italie et Méditerranée modernes et contemporaines*, fasc. 133 (2021): 159–72. https://doi.org/10.4000/mefrim.10674.

[6] Chiaravalloti, Maria Teresa, e Maria Taverniti. «Strumenti e terapie nelle cure psichiatriche: estratti dall'archivio storico dell'ospedale psichiatrico di Girifalco». *AIDA INFORMAZIONI* 1–2 (2017): 33–47.

[7] Chiaravalloti, Maria Teresa, Maria Taverniti, e Maria Francesca Dovetto. «Le cartelle dell'ex ospedale psichiatrico di Girifalco. Lessico, strumenti e terapie». In *Lingua e patologia. I sistemi instabili*, a cura di Maria Francesca Dovetto, 235–68. Roma: Aracne Editrice, 2020.

[8] Codepé, Maurizio. *La carta e il suo degrado*. Arte e restauro Strumenti 3. Firenze: Nardini Editore, 1991.

[9] Cupidi, Chiara, Valentina Laganà, Nicoletta Smirne, e Amalia Cecilia Bruni. «The Role of Historical Medical Archives in the Genealogical Rebuilding of Large Families Affected by Neurodegenerative Diseases». *Journal of Neurology & Neuromedicine* 2, fasc. 5 (15 maggio 2017). https://doi.org/10.29245/2572.942X/2017/5.1127.

[10] Dovetto, Francesca Maria. «I marginali dell'ex Ospedale psichiatrico di Girifalco e il lessico delle malattie di nerve alla testa». In *In limine. Frontiere e integrazioni*, a cura di Diego Poli, 137–61. Roma: Il Calamo, 2019.

[11] Dovetto, Francesca Maria. «Terminologia 'calda' e terminologia 'fredda': alcune caratteristiche della costituzione del lessico italiano della fonetica». In *Le parole per le parole. I Logonimi nelle lingue e nel metalinguaggio*, a cura di Cristina Vallini, 279–300. Roma: Il Calamo, 2000.

[12] Fianco, Renato. *L'asilo della maggior sventura. Origini e sviluppo del manicomio veronese di San Giacomo di Tomba (1880-1905).* Verona: Cierre edizioni, 1992.

[13] Fiorino, Vinzia. *Matti, indemoniate e vagabondi: dinamiche di internamento manicomiale tra Otto e Novecento*. Venezia: Saggi Marsilio, 2002.

[14] Gaino, Alberto. *Il manicomio dei bambini: storie di istituzionalizzazione*. Torino: Edizioni Gruppo Abele, 2017.

[15] Greco, Oscar. *I demoni del Mezzogiorno: follia, pregiudizio e marginalità nel manicomio di Girifalco (1881-1921).* Soveria Mannelli: Rubbettino, 2018.

[16] Khurana, Diksha, Aditya Koli, Kiran Khatter, e Sukhdev Singh. «Natural Language Processing: state of the art, current trends and challenges». *Multimedia Tools and Applications* 82 (2023): 3713–44. https://doi.org/10.1007/s11042-022-13428-4.

[17] Lagonia, Paolo. «Storia di un luogo di ordinarie follie». *Calabria Produttiva*, 2009.

[18] Lagonia, Paolo, Matteo Aloi, Fabio Magliocco, Gregorio Cerminara, Cristina Segura-Garcia, Valeria Del Vecchio, Mario Luciano, Andrea Fiorillo, e Pasquale De Fazio. «First World War and Mental Health: a retrospective comparative study of veterans admitted to a psychiatric hospital between 1915 and 1918». *Riv Psichiatr* 52, fasc. 3 (2017): 120–25. https://doi.org/10.1708/2722.27764.

[19] Li, Irene, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, et al. «Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review». *arXiv*, 6 luglio 2021. https://arxiv.org/abs/2107.02975.

[20] Marcello, Domenico. *Un secolo di manicomio: storia del Manicomio di Girifalco*. Catanzaro: Vincenzo Ursini Editore, 1995.

[21] Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, et al. «Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study». *Journal of Documentation* 75, fasc. 5 (settembre 2019): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

[22] Onan, Aytuğ, Serdar Korukoğlu, e Hasan Bulut. «Ensemble of keyword extraction methods and classifiers in text classification». *Expert Systems with Applications* 57 (15 settembre 2016): 232–47. https://doi.org/10.1016/j.eswa.2016.03.045.

[23] Pasceri, Erika, Mérième Bouhandi, Claudia Lanza, Anna Perri, Valentina Laganà, Raffaele Maletta, Raffaele Di Lorenzo, e Amalia C. Bruni. «Neurodegenerative clinical records analyzer: detection of recurrent patterns within clinical records towards the identification of typical signs of neurodegenerative disease history». *JLIS.It* 14, fasc. 2 (2023): 20–38. https://doi.org/10.36253/jlis.it-522.

[24] Roshdi, Akram, e Akram Roohparvar. «Review: Information Retrieval Techniques and Applications». *International Journal of Computer Networks and Communications Security* 3, fasc. 9 (2015): 373–77.

[25] Valeriano, Annacarla. Malacarne: donne e manicomio nell'Italia fascista. Roma: Donzelli, 2018.

[26] Vanzan Marchini, Nelli Elena. «L'eterogeneità dei patrimoni ospedalieri: i problemi della conservazione e della valorizzazione». In *Medicina e ospedali. Memoria e futuro: aspetti e problemi degli archivi sanitari. Atti del Convegno, Napoli, 20-21 dicembre 1996*, 41–52. Roma: Ministero per i beni e le attività culturali, Direzione generale archivi, 2001.

Preservazione del patrimonio culturale e cloud computing: caratteristiche e criticità

Manuela Grillo
Sapienza Università di Roma, Italia – manuela.grillo@gmail.com

ABSTRACT[1]

Il contributo intende analizzare caratteristiche e criticità dei servizi di *cloud computing* per la preservazione del patrimonio culturale: l'uso consapevole della tecnologia, orientato al principio della condivisione e della cittadinanza globale, passa anche dalla conoscenza delle caratteristiche tecniche e contrattuali dei servizi di cloud computing, nonché dalla consapevolezza delle criticità (es. sicurezza dei dati, ecc.). Attualmente la preservazione dei dati digitali del patrimonio culturale - così come l'intero processo di digitalizzazione della Pubblica Amministrazione in genere - vede il *cloud computing* come strumento cardine, secondo il paradigma "Cloud first" definito dal Piano nazionale di ripresa e resilienza (PNRR). Le istituzioni della memoria - archivi e biblioteche in particolar modo – hanno da sempre avuto il ruolo di garantire accesso ai beni culturali anche nel lungo periodo, per cui sono direttamente coinvolte nel dibattito in corso su metodi e strumenti di conservazione ed accesso permanente al patrimonio culturale in formato digitale.

PAROLE CHIAVE

Cloud computing; Data sovereignty; Cultural Heritage.

1. INTRODUZIONE

Le istituzioni della memoria - archivi e biblioteche in particolar modo - hanno da sempre avuto il ruolo di garantire accesso ai beni culturali anche nel lungo periodo. Attualmente la preservazione dei dati digitali del patrimonio culturale - così come l'intero processo di digitalizzazione della Pubblica Amministrazione in genere - vede il *cloud computing* come strumento cardine, secondo il paradigma "Cloud first" definito dal Piano nazionale di ripresa e resilienza (PNRR). In riferimento al documento dell'Agenzia per l'Italia digitale "Progetto Poli di conservazione: definizione di un modello di riferimento per i Poli di Conservazione e della relativa rete nazionale"[2], nel luglio del 2021 l'Associazione Italiana Biblioteche ha inviato ad AgID una puntuale comunicazione[3], sottolineando come la questione della conservazione nel lungo periodo di beni culturali in formato digitale - questione che riguarda molto direttamente le biblioteche, in particolare le due Biblioteche nazionali centrali di Firenze e Roma - non sia più rinviabile. Nel documento presentato da AIB in occasione della presentazione delle Osservazioni al PNRR[4] AIB propone la creazione di una società pubblica per la gestione dei servizi di conservazione e accesso permanente al patrimonio culturale.

La questione va ben oltre l'uso delle tecnologie di cloud computing per la gestione delle biblioteche e delle loro collezioni [8] e richiama il dibattuto tema della sovranità digitale (*data sovereignty)* [9, 11]. Le criticità legate all'outsourcing per la preservazione dei dati digitali – in termini di perdita di controllo, sia sulla gestione del dato che sui livelli di sicurezza – sono da tempo note: nonostante le grandi organizzazioni possiedano un potere negoziale maggiore rispetto ai singoli utenti, in virtù del quale possono ottenere delle deroghe sui contratti standard, permane comunque una generalizzata asimmetria contrattuale tra i Cloud Provider e gli utenti finali, caratterizzata da una rigidità e complessità dei termini di contratto, da clausole fortemente sbilanciate a favore dei fornitori e dalla mancanza di trasparenza sull'erogazione del servizio (disponibilità, misure di sicurezza, dislocazione dei data server, policy di backup, ecc.) [7, 3].

Non soltanto privati cittadini ma soprattutto organizzazioni, sia private che pubbliche, usufruiscono di servizi cloud e affidano ad una logica di outsourcing la gestione del proprio sistema informativo e culturale: informazioni riguardanti ministeri, enti di ricerca, comuni, regioni, dati sull'ordine pubblico, sulla salute pubblica, nonché quelli relativi all'intero patrimonio culturale (se si considera non solo il patrimonio culturale nativo digitale ma anche tutta la mole di dati derivante dai progetti di digitalizzazione dell'analogico), quindi, in definitiva, i dati strategici di un intero Paese vengono custoditi

[1] Ringrazio l'Ing. Daniele Ranzino per i chiarimenti tecnici e la messa a disposizione dei materiali.
[2] https://www.agid.gov.it/it/piattaforme/conservazione/poli-conservazione
[3] https://www.aib.it/notizie/conservazione-accesso-permanenti-patrimonio-digitale/
[4] https://www.aib.it/notizie/osservazioni-aib-recovery-plan-piano-ripresa-resilienza/

in cloud da compagnie straniere, legate a doppio filo ai rispettivi governi (basti pensare agli esiti del Progetto Google Books[5]).

2. CLOUD COMPUTING: ELEMENTI CARATTERIZZANTI

Secondo la definizione del National Institute of Standards and Technology (NIST)[6], il *cloud computing* è un modello di infrastrutture informatiche che permette un facile accesso da remoto a risorse computazionali condivise e configurabili (ad esempio reti, server, storage, applicazioni, servizi) a richiesta, in qualsiasi momento e in autonomia, senza necessità di interazione umana da parte del service provider (*on demand* e *self service*); l'accesso alle risorse deve avvenire tramite la rete attraverso qualsiasi tipologia di *device* (pc, tablet, smartphone, etc.)[7].

Le risorse devono inoltre rispettare le seguenti caratteristiche:

- *Resource Pooling*: le risorse del service provider sono messe in comune per servire molteplici utenti, attraverso un modello condiviso, e assegnate dinamicamente in base alle richieste dell'utente;
- *Rapid Elasticity*: le risorse possono essere assegnate e rilasciate rapidamente ed elasticamente, in alcuni casi anche automaticamente, per scalare velocemente le capacità del sistema in relazione alla domanda;
- *Measured service*: l'utilizzo delle risorse viene costantemente monitorato e ottimizzato ad un livello di astrazione dipendente dalla tipologia di servizio offerto, come memoria, elaborazione, larghezza di banda, utenti attivi, ecc.

Il NIST ha inoltre introdotto tre modelli di servizio:

- Iaas (*Infrastructure as a Service*): il cloud service provider consente all'utente di creare la propria infrastruttura informatica virtuale in base alle proprie necessità, secondo un modello di servizio pay as you go (pagamento a consumo), gestendo direttamente risorse di calcolo, storage, risorse di rete, server e la configurazione dei sistemi facenti parte dell'infrastruttura (Sistemi operativi, applicazioni, dati, ecc.).
- PaaS (*Platform as a Service*): l'utente predispone un ambiente per lo sviluppo, l'esecuzione e la gestione di proprie applicazioni, all'interno di un ambiente e un'infrastruttura fornita dal cloud provider.
- SaaS (*Software as a Service*): il cloud provider mette a disposizione dell'utente servizi e software accessibili via rete tramite diverse tipologie di device.

Lo strato logico e fisico sottostante a ogni modello di servizio è completamente trasparente - "trasparente" in senso informatico, ovvero non visibile e comunque non importante - per l'utente, il quale detiene sempre meno controllo e conoscenza dell'infrastruttura hardware e software man mano che ci si allontana da un modello di tipo IaaS e ci si avvicina a soluzioni di tipo SaaS. In ogni caso, spesso la differenza tra i diversi modelli di servizio non è netta, pertanto negli ultimi anni si fa riferimento a un modello di servizio più generico denominato XaaS (*Anything as a Service*).

Le tecnologie cloud si articolano secondo quattro modelli di distribuzione: cloud pubblico (modalità più diffusa di tecnologia cloud, in cui il cloud provider, che possiede l'infrastruttura, fornisce servizi cloud ad una molteplicità di utenti finali); cloud privato (o *on premise*, modalità in cui risorse infrastrutturali vengono gestite e utilizzate in sede da una singola organizzazione o esternamente da terze parti); community cloud (le risorse infrastrutturali vengono condivise da un gruppo di organizzazioni con esigenze e missioni comuni); cloud ibrido (l'infrastruttura informatica è implementata combinando diverse tipologie di cloud in parte *on premise* e in parte in cloud (pubblico o community) [2: 65-94].

Con l'affermarsi del cloud computing, i contenuti digitali sono passati dall'essere circoscritti in uno spazio fisico ben definito all'essere accessibili e disponibili da ogni luogo (e potenzialmente da chiunque) attraverso una moltitudine di dispositivi, mediante i quali è possibile accedere ai servizi sempre disponibili. Non essendo più necessario possedere risorse, con il cloud computing si afferma quindi il concetto di fruizione di servizi rappresentato dal paradigma aaS (as a service) che avviene su richiesta (*on demand*) adattandosi al reale fabbisogno dell'utente finale attraverso un'ottimizzazione delle risorse e, in definitiva, un abbattimento dei costi rispetto all'insourcing.

3. IMPATTO DEI SERVIZI DI CLOUD COMPUTING SULLE ISTITUZIONI CULTURALI

Se da un lato le soluzioni di cloud computing comportano anche per le organizzazioni culturali un enorme risparmio in termini di investimento tecnologico hardware e software (acquisto, installazione, gestione, manutenzione, smaltimento del sistema informativo, compresi i costi di formazione di personale qualificato dedicato alla sua gestione), dall'altro presentano anche forti elementi di criticità in grado di minare la riservatezza, l'integrità e la disponibilità dei dati.

[5] Sul progetto si vedano le informazioni di avvio in [1] e i dettagli operativi e le valutazioni costi/benefici in [6]; entrambi gli autori dei contributi sono stati Direttori della Biblioteca Nazionale Centrale di Roma, biblioteca coordinatrice esecutiva dell'intero progetto.
[6] https://www.nist.gov/
[7] https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

In primis vi è una problematica legata alla disponibilità e alla continuità operativa dei servizi erogati da parte dei Cloud Service Provider: potrebbe capitare, per quanto raro, che i servizi offerti dal cloud provider possano essere vittima di incidenti e risultare indisponibili, compromettendo la fruibilità del patrimonio culturale. Così come una ulteriore problematica è rappresentata dalle dinamiche di *vendor lock-in*, ovvero la situazione di dipendenza da un unico fornitore, che non può essere sostituito senza gravi conseguenze (il che costituisce una condizione particolarmente fastidiosa nel caso in cui ci si ritrovi legati contrattualmente a un cloud provider poco flessibile).
Attraverso il cloud computing di fatto le istituzioni culturali consegnano il proprio patrimonio ai cloud provider, perdendone il controllo diretto ed esclusivo. Inoltre, in un contesto cloud, una moltitudine di soggetti possono essere coinvolti nel ciclo di vita del dato (acquisizione, trattamento, archiviazione, backup, ecc.): un fornitore di servizi cloud può infatti delegare a terze parti la gestione dell'infrastruttura o parte di essa. Questa possibilità di collaborazione tra diversi cloud provider, in maniera totalmente trasparente per l'utente finale, pone diversi interrogativi.
Non ultime, soprattutto per il patrimonio archivistico, le questioni legate alla gestione dei dati personali: venendo meno il controllo sul ciclo di vita del dato, le policy di storage e di backup, non è chiaro come i dati personali vengono trattati dai service provider. Non solo l'utente istituzionale talvolta ignora l'esatta geolocalizzazione dei server che custodiscono i propri dati, ma persino per gli stessi cloud provider non è scontato sapere esattamente dove risiedono i dati, a causa della crescente complessità architetturale delle infrastrutture e del coinvolgimento a vario titolo di terze parti (altri cloud provider, data broker, compagnie ad-tech, ecc.). Nonostante le tutele introdotte in ambito europeo dal Regolamento generale sulla protezione dei dati (General Data Protection Regulation, GDPR)[8], i dati, in una frazione di secondo, possono fare il giro del mondo ed essere distribuiti ad una moltitudine di server residenti in diversi Paesi, nei quali vigono ordinamenti differenti, il tutto in assenza di controllo sul ciclo di vita del dato, sulle logiche e policies di trattamento [4].
Per ciò che riguarda le criticità create da ordinamenti giuridici differenti, basta pensare al contrasto tra GDPR e Cloud Act[9], che permette alle autorità statunitensi, alle forze dell'ordine e alle agenzie di intelligence di accedere ai dati digitali gestiti da operatori di servizi di cloud computing, a prescindere dal luogo in cui questi dati si trovano, anche su server fuori dal territorio degli Stati Uniti, quando gli operatori sono sottoposti alla giurisdizione statunitense o anche quando gli operatori hanno una filiale nel territorio USA o operano nel mercato americano.
Chiaramente i cloud service provider devono conformarsi ai principi tutelati dagli istituti contrattuali [5], ma le differenze culturali stesse tra Europa e Stati Uniti – i maggiori cloud provider sono statunitensi – sul concetto di privacy collidono con le prescrizioni del GDPR. Un emblematico caso di collisione è infatti la differenza di approccio al principio stesso di privacy tra un sistema di tipo common law come quello americano e un sistema giuridico come quello europeo: negli Usa infatti la privacy è considerata alla stregua di un diritto commerciale, a prevalente aspetto economico, mentre in Europa invece viene considerata un diritto fondamentale della persona; per questo motivo, ad esempio, a livello europeo viene tutelato il diritto all'oblio, mentre negli Stati Uniti è percepito come un vulnus alla libertà d'espressione.
In un contesto come quello attuale, in cui più del 70% del mercato cloud è appannaggio di società americane (Amazon, Microsoft, Google, ecc.) questi aspetti sono tutt'altro che secondari, pur eliminando qualunque approccio formale e/o legato ad aspetti idealizzanti filosofico-normativi.
La perdita di sovranità in ambito tecnologico mina i principi di indipendenza ed autodeterminazione di interi Paesi ed è in questo contesto che si colloca il dibattito sulla necessità di creazione di uno spazio immateriale in cui poter affermare la "sovranità digitale" dei Paesi europei, ovvero la necessità di un cloud nazionale e/o europeo, realtà ancora lontana dalla sua concreta realizzazione.
In questo senso, già nel 2019, Antonello Soro - Presidente dell'autorità amministrativa indipendente Garante per la protezione dei dati personali, dal 19 giugno 2012 al 28 luglio 2020 - durante la presentazione al Parlamento della Relazione annuale 2019 affermava: «la stretta dipendenza della sicurezza della rete da chi ne gestisca i vari snodi e 'canali' induca a ripensare il concetto di sovranità digitale. E di fronte alla delocalizzazione in cloud di attività rilevantissime chiediamo al Parlamento e al Governo se non si debba investire in un'infrastruttura cloud pubblica, con stringenti requisiti di protezione, per riversarvi con adeguata sicurezza dati di tale importanza. In un contesto in cui le tecnologie ICT sono divenute - sempre più chiaramente con la pandemia – la principale infrastruttura di ciascun Paese, assicurarne una regolazione sostenibile e adeguata, tale da garantire sicurezza, indipendenza dai poteri privati, soggezione alla giurisdizione interna, diviene un obiettivo non più eludibile»[10].

[8] Il testo del Regolamento (UE) 2016/679 (aggiornato alle rettifiche pubblicate sulla Gazzetta Ufficiale dell'Unione europea 127 del 23 maggio 2018) è disponibile sul sito del Garante per la protezione dei dati personali (https://www.garanteprivacy.it/il-testo-del-regolamento), in una versione "arricchita" del testo per offrire una lettura più ampia e ragionata della nuova normativa.
[9] https://www.congress.gov/bill/115th-congress/house-bill/4943
[10] https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9428236

Al momento attuale, il relativamente recente lancio[11] da parte della Commissione europea dell'European Collaborative Cloud for Cultural Heritage - seguito a livello nazionale da APRE (Agenzia per la Promozione della Ricerca Europea)[12] - offre delle nuove prospettive: l'European Collaborative Cloud for Cultural Heritage sarà un'infrastruttura digitale che consentirà una collaborazione transdisciplinare e su larga scala tra i professionisti del settore in tutta l'UE, attraverso strumenti digitali all'avanguardia. Il budget previsto da Horizon Europe, di 110 milioni di euro fino al 2025, è dedicato a sostenerne lo sviluppo, con l'intento di creare un cloud che promuova la cooperazione e la co-creazione tra i settori culturali, creativi e tecnologici e che contribuisca a salvaguardare i tesori culturali europei attraverso la formazione di un'infrastruttura digitale. L'infrastruttura è prevista come fornitrice di tecnologie all'avanguardia per la digitalizzazione di beni, per la ricerca di opere d'arte e per la raccolta di dati. Considerato che al momento non è disponibile[13] che il "Report on a European Collaborative Cloud for Cultural Heritage"[14], stilato dagli otto esperti incaricati della valutazione, la speranza è che le sorti dell'infrastruttura siano diverse da quelle dell'iniziativa europea Gaia-X[15]: il progetto ha rappresentato il primo passo verso una strategia comune di infrastruttura dati europea, alla base del quale vi è il principio di data sovereignty e la necessità di indipendenza dalle big tech per l'archiviazione e la condivisione di informazioni sensibili. L'obiettivo del progetto non è la creazione di un singolo Cloud centralizzato, ma piuttosto di un ecosistema, un'infrastruttura dati federata (formata da diversi Cloud Provider che operano secondo standard e principi comuni), trasparente, affidabile, interoperabile e sicura, in cui i dati vengano condivisi e resi disponibili, restituendone il controllo e la sovranità agli utenti. L'introduzione di regole comuni tra i servizi Cloud all'interno dell'Unione Europea permette a Gaia-X di armonizzare la condivisione dei dati all'interno di un ecosistema federato, aperto e interoperabile, basato sul rispetto delle normative europee.

La struttura organizzativa di Gaia-X si articola in tre componenti principali:

- l'associazione Gaia-X (Gaia-X European Association for Data and Cloud): fulcro della struttura organizzativa, è un'associazione senza scopo di lucro, nata nel gennaio 2021 da 22 organizzazioni pubbliche e private (11 tedesche e 11 francesi), ad oggi costituita da più di 340 membri tra aziende, istituti di ricerca, associazioni, pubbliche amministrazioni e istituzioni politiche;
- gli Hub che fungono da collegamento tra le esigenze locali e l'associazione Gaia-X. Identificano le esigenze di un territorio sviluppando soluzioni tecnologiche compatibili con il framework, anche attraverso la collaborazione di altri Hub;
- la Community: gli stakeholders contribuiscono attivamente ai gruppi di lavoro, condividendo le loro conoscenze tramite piattaforme di collaborazione, eventi, conferenze, workshop e webinar.

Il framework si basa sul concetto di decentralizzazione, per cui ogni fornitore di servizi opera in completa autonomia, recependo però gli standard Gaia-X e garantendo sicurezza e interoperabilità con gli altri fornitori coinvolti.

La forte spinta verso soluzioni open source è il vero punto di forza del progetto, ma se l'obiettivo originale era quello di guidare la data strategy europea, all'interno di un'Europa che eserciti pienamente la propria sovranità digitale, al netto della retorica del lancio del progetto, va purtroppo notato che dell'obiettivo originale rimane ben poco: sono infatti parte integrante del progetto Amazon, Microsoft, Google e Alibaba, ossia quegli stessi big player il cui predominio doveva essere controbilanciato proprio con il progetto Gaia-X. Anche l'ingresso di multinazionali dei dati come Oracle e Salesforce, delle telecomunicazioni come Huawei, insieme a colossi dell'informatica come Hp e Ibm, fanno sorgere dei dubbi sulla direzione e sulla forma che sta prendendo il progetto, così come la stessa diffusione degli Hub Gaia-X al di fuori del territorio europeo – con tutte le implicazioni giuridiche a questo connesse - rappresentano un forte campanello d'allarme.

4. CONCLUSIONI

Il ritardo italiano ed europeo in genere nell'ambito delle ICT e la non pienamente consapevole introduzione delle stesse tecnologie - in particolare delle tecnologie di cloud computing - hanno accelerato il processo di perdita di sovranità su dati, servizi e tecnologie. La perdita di sovranità si manifesta innanzitutto nella legittimazione del ruolo dei privati nella cogestione di settori strategici quali la sicurezza nazionale, l'attività giudiziaria, la sanità, la ricerca, così come la conservazione del patrimonio culturale. Ciò può tradursi in inevitabili ingerenze straniere sulle scelte politiche nazionali e comunitarie che ne minacciano la sovranità [10]. Al mutare degli assetti geopolitici – ad esempio in caso di crisi diplomatica o anche in un contesto di guerra commerciale e/o militare – i servizi Cloud stessi alla base di asset strategici potrebbero essere resi indisponibili e smettere di funzionare.

[11] https://digitallibrary.cultura.gov.it/eventi/un-cloud-per-tutti/
[12] https://apre.it/a-cloud-for-all-linfrastruttura-digitale-per-il-patrimonio-culturale/
[13] https://ec.europa.eu/commission/presscorner/detail/en/IP_22_3855
[14] https://op.europa.eu/en/publication-detail/-/publication/90f1ee85-ca88-11ec-b6f4-01aa75ed71a1/language-en
[15] https://gaia-x.eu/

La complessità e le sfide poste dalle nuove tecnologie digitali rendono non semplice individuare ed attuare azioni di salvaguardia istituzionale attraverso istituti contrattuali: del resto come può attuarsi una vera sovranità digitale e dei dati senza un controllo diretto delle infrastrutture tecnologiche? Circa il rispetto delle normative nazionali o europee, si hanno o si potrebbero avere reali garanzie da parte di multinazionali che possiedono un potere analogo, se non superiore, a quello degli Stati nazionali? Impossibile non condividere l'affermazione dell'ex Presidente del Garante per la protezione dei dati personali, Antonello Soro, secondo cui «[…] i gestori delle grandi piattaforme tecnologiche hanno profondamente cambiato la geografia del potere globale, con straordinaria capacità di condizionamento delle politiche pubbliche oltre che del comportamento individuale» [12: 19].

Le criticità esposte pongono forti preoccupazioni e offrono elementi da attenzionare per un utilizzo consapevole delle tecnologie di cloud computing. È necessario che la praticità offerta dalla tecnologia si coniughi alla protezione dei dati e del patrimonio culturale, così come degli interessi degli utenti istituzionali della pubblica amministrazione e della ricerca pubblica, che come espressioni della nostra identità vanno tutelati, al di là dei giganti interessi economici sottesi alla fornitura di soluzioni ICT.

BIBLIOGRAFIA

[1] Avallone, Osvaldo. 'Il Progetto Google Books: La prima grande esperienza di accesso diretto al patrimonio bibliografico nazionale'. *DigItalia* 8, no. 1 (2013): 9–13. https://digitalia.cultura.gov.it/article/view/716.

[2] Bellini, Francesco, e Fabrizio D'Acenzo, (a cura di). *Digital Transformation and Data Management.* Pisa: Pacini giuridica, 2020.

[3] Bellotti, Roberto, (a cura di). *Il Cloud Computing nelle imprese e nella Pubblica Amministrazione*. Milano: Giuffrè Francis Lefebvre, 2019.

[4] Biasiotti, Adalberto. *Il nuovo regolamento europeo sulla protezione dei dati: una guida pratica alla nuova privacy e ai principali adempimenti del regolamento UE 2016/679, Aggiornata al D.Lgs. 101/2018*. Roma: EPC, 2018.

[5] Boncinelli, Vanni. 'Modelli tecnici e disciplina giuridica del c.d. Cloud Computing'. *Rivista Italiana Di Informatica e Diritto* 3 (2021): 27–45. https://doi.org/10.32091/RIID0023.

[6] De Pasquale, Andrea. 'L'attuazione in Italia del progetto GoogleBooks'. *DigItalia* 14, no. 1 (2019): 103–13. https://digitalia.cultura.gov.it/article/view/2277.

[7] Faggioli, Gabriele, and Annamaria Italiano. *I contratti di Cloud Computing. Comprendere, affrontare e negoziare i contratti con i provider*. Milano: Franco Angeli, 2017.

[8] Kipps, Kayla, e Allison Kaiser Jones. *Collection Management in the Cloud: A Guide for Using Cloud Computing Technologies in Libraries.* Lanham: Rowman & Littlefield, 2022.

[9] Lukings, Melissa, e Arash Habibi Lashkari. *Understanding Cybersecurity Law in Data Sovereignty and Digital Governance: An Overview from a Legal Perspective*. Cham: Springer, 2022.

[10] Monti, Andrea. 'Scienza, Tecnocontrollo e Public-Policy Nell'era COVID-19'. *Rivista Trimestrale Di Scienze'amministrazione* 2 (2020): 1–37. https://rtsa.eu/RTSA_2_2020_Monti.pdf.

[11] Paganelli, Valentina. 'Conservazione dei dati e sovranità digitale. Una rilettura della (Big) Data Governance Pubblica alla luce delle nuove sfide globali'. *Rivista Italiana Di Informatica e Diritto* 3 (2021): 11–26. https://doi.org/10.32091/RIID0022.

[12] Soro, Antonello. *Democrazia e potere dei dati: Libertà, algoritmi, umanesimo digitale*. Milano: Baldini&Castoldi, 2019.

Preserving Culinary Traditions. A Crowdsourced Digital Collection of Cookbooks

Giulia Renda[1], Giulia Manganelli[2], Mila Fumini[3], Marilena Daquino[4]
[1] University of Bologna, Italy - giulia.renda3@unibo.it
[2] Independent researcher, Italy - giuliamanganelli@me.com
[3] Independent researcher, Italy - mila.fumini@me.com
[4] University of Bologna, Italy - marilena.daquino2@unibo.it

ABSTRACT
Recipes of popular origin and handwritten cookbooks are often overlooked by scholars. *Ragù* is a pilot project that tries to fill in this gap by gathering and digitising a collection of cookbooks belonging to the Italian traditional cuisine, and making it accessible via a digital platform. The project aims at contributing to two research lines: a) to identify agile methods for publishing data in a low-cost crowdsourcing project, and b) to devise an effective storytelling journey for the *Ragù* project.

KEYWORDS
Cultural Heritage preservation; Crowdsourcing; Data management; Digital storytelling.

1. INTRODUCTION
Recipes of popular origin and handwritten cookbooks are often overlooked by scholars, despite representing a rich source of cultural insight, since they present the influence of diverse cultures and their relationship with traditions, and witness the evolution of material history. Yet, extensive projects to collect, transcribe, and preserve such a precious intangible cultural heritage are missing. *Ragù* is a pilot project that tries to fill in this gap by collecting and digitising a collection of cookbooks belonging to the Italian traditional cuisine, and making it accessible via a digital platform.
Being a low-budget project managed by a few volunteers, the project presents several criticalities. Firstly, the primary sources need to be physically collected, transcribed and digitised and a mid-term data management plan for preservation and exploitation is needed. Additionally, mechanisms for effective metadata extraction and query must be designed to populate user-friendly interfaces. Given the resource constraints of the project, such methods must ensure easy updates (to both metadata and user interfaces) and access to data and images. Lastly, since the web platform aims to engage both everyday users open to serendipitous discovery and scholars interested in exploring data with more sophisticated queries, appropriate strategies for dissemination must be designed.
In this article, we present the strategies adopted in the *Ragù* project. We describe existing approaches to crowdsourcing and storytelling (section 2), we propose a simplified model for data management in low-cost projects (section 3), and we present a digital storytelling journey for the case study at hand (section 4).

2. STATE OF THE ART
In the last decades, the massive digitisation process of large libraries, museums and archives has been successfully carried out by specialists in galleries, libraries, archives and museums (GLAM) [9]. However, the business model applied by national and international institutions is not always applicable to small and medium-sized institutions, cultural associations or individuals, who may have to find alternative strategies.
Crowdsourcing [6] is the process of outsourcing tasks or problems to a large group of people, typically using the web as a platform for information exchange. Crowdsourcing methods have been adopted in GLAM and heritage organisations to carry out several tasks, such as metadata creation, analysis of cultural heritage objects, as well as contributions of private objects or experiences [13, 1, 10]. Notably, public collections rarely contain materials belonging to private collections or individuals, e.g. household documents. Therefore crowdsourcing such primary sources may be the only way to foster research enquiries [14] based on witnesses of our material history. An exemplar case is the Europeana 1914-1918[1] campaign, where users could upload pictures of memorabilia to be digitised and populate a digital archive.
However, similar projects on a smaller scale in terms of both finances and workforce encounter problems related to continuity and sustainability [3], including socio-technical, technological, and dissemination challenges. As a matter of fact, most crowdsourcing projects in GLAM rely on expensive and not easy-to-customise platforms for collecting citizens' contributions [3].

[1] https://www.europeana.eu/en/collections/topic/83-world-war-i

More importantly, existing projects do not often provide dissemination interfaces to communicate results of the crowdsourcing campaign and may not be able to engage a wide variety of visitors. From the literature on approaches for information seeking purposes, two types of interactive behaviour can be identified. The first strategy focuses on reducing the cognitive load of users by breaking down information, displaying it in small chunks [7]. The second approach advocates for more *generous interfaces*, arguing that hiding the entirety of content from the user can become a source of frustration [16]. Considering that visitors on cultural heritage websites don't always have a specific task in mind, methods for casual browsing can effectively help them to discover and find their way to a specific task or interest. Such an approach has been summarised as "overview first, zoom and filter, then details on demand" [12]. To overcome the overwhelming feeling caused by the large amount of materials that a digital collection can contain, it is common practice in the cultural heritage domain to design "paths" to help the user navigate content, and to use data visualisation techniques to provide overviews and improve their experience [4]. However, data visualisations can be hard to understand for everyday users. To mitigate this complexity, storytelling techniques can be used to convey complex information compactly [5, 8]. The telling of a story (path) acts as an interpretative framework [2].

Examining projects that use storytelling and data visualisation as tools for presenting digital collections of cookbooks and guiding the user exploration, we could find very few examples. *The Early American Cookbooks* [15] collects cookbooks from 1800 to 1920, and serves them on a website organised in thematic categories, which feature blog-like articles using data visualisations to offer insights into the collection. Developed by a digital humanist at the New York University Libraries, the workflow for data management does not seem to be replicable, and the storytelling strategies are limited to few reiterated charts that do not offer means to continue the exploration in depth. Other examples of collections of private cookbooks include *The South African Jewish Cookbook Project*[2], but the project website does not present the collection using data visualisations or storytelling strategies. Rather it offers traditional interfaces for exploring items of the collection thematically. Therefore, this article aims to contribute to two lines of research: a) to identify agile methods for publishing data in a low-cost project (section 3), and b) to devise an effective storytelling journey for crowdsourced cookbooks collections (section 4).

3. AN OPEN SOURCE AND REUSABLE APPROACH TO MANAGE LOW-COST CROWDSOURCING CAMPAIGNS

The project began with a serendipitous discovery of handwritten cookbooks in a cellar to be cleared out. Historian Mila Fumini, leading the project, initiated the first call for contributions in 2019. Collection efforts have mainly occurred through public appeals at food-related events or locations, with additional outreach via email, followed by visits to homes to assess potential materials. The only criterion for inclusion in the collection is that notebooks must be handwritten. Whenever a new cookbook is proposed and evaluated, it is photographed with a smartphone and then returned to the owner. Pictures[3] are archived using a private online folder, and selected contents are transcribed in a shared spreadsheet online. The table was devised by a restricted group of data curators that volunteered in the project and it includes bibliographic metadata of cookbooks and recipes - e.g. recipe title, year, city of origin, author - information on the image file, and other insights relevant to historical research enquiries, e.g. type of recipe, ingredients, quantities, and linguistic variations of names (see Fig. 1). Controlled vocabularies and folksonomies have been defined for ingredients and categories, geographical information (city, region, and country), scope, procedure, and units of measurement. Future plans include aligning with existing resources. The working table is available as a Google Spreadsheet[4] and can be downloaded in CSV format for analysis purposes.

	B	C	D	E	F	G	H	I	J	K	L	M
1	Notebook	From	To	Time qualifier	Place	Region	Country	Author	Img	Pag.	Title chapter	Title Recipe
2	title	year	year		town			surname name	nome	numero		
8	Le ricette di zia Dina	1960	1970	ca.	Rimini	Emilia R...	Italy	Dina	Quaderno 1_Rimini_29ago2019_2.jpg	2	Minestre	Maccheroni con besciamella
9	Le ricette di zia Dina	1960	1970	ca.	Rimini	Emilia R...	Italy	Dina	Quaderno 1_Rimini_29ago2019_2.jpg	2	Minestre	Maccheroni con besciamella
10	Le ricette di zia Dina	1960	1970	ca.	Rimini	Emilia R...	Italy	Dina	Quaderno 1_Rimini_29ago2019_2.jpg	2	Minestre	Maccheroni con besciamella
11	Le ricette di zia Dina	1960	1970	ca.	Rimini	Emilia R...	Italy	Dina	Quaderno 1_Rimini_29ago2019_2.jpg	2	Minestre	Maccheroni con besciamella
12	Le ricette di zia Dina	1960	1970	ca.	Rimini	Emilia R...	Italy	Dina	Quaderno 1_Rimini_29ago2019_2.jpg	2	Minestre	Maccheroni con besciamella
13	Le ricette di zia Dina	1960	1970	ca.	Rimini	Emilia R...	Italy	Dina	Quaderno 1_Rimini_29ago2019_2.jpg; Quaderno 1_Rimini_29ago2019_3.jpg	2; 3	Minestre	Polenta alla lombarda
14	Le ricette di zia Dina	1960	1970	ca.	Rimini	Emilia R...	Italy	Dina	Quaderno 1_Rimini_29ago2019_2.jpg; Quaderno 1_Rimini_29ago2019_3.jpg	2; 3	Minestre	Polenta alla lombarda

Figure 1. An extract from the Google Spreadsheet

[2] https://sajewishcookbooks.org.za/
[3] Available under a CC-BY 4.0 licence.
[4] https://docs.google.com/spreadsheets/d/1terFx_mYVspOjvDUJfcBqHR7j_OlvXVwUesHNHAl_wU/edit?usp=sharing

Digitisation and transcription is an ongoing process that began in 2023, with currently digitised materials spanning approximately 460 recipes from the 1960s to the 1980s. To showcase (partial) results of the data collection as soon as possible, we decided to create a website that could be easily and dynamically updated with new recipes. To achieve this, we leverage one of the most well-known platforms for depositing open source code, i.e. GitHub. GitHub provides the *Ragù* project with four important features, namely: (1) free hosting of data dumps, i.e. a versioned copy of the aforementioned working table, (2) free-of-charge hosting and publication of a static website, (3) editorial strategies to prevent people to immediately publish content on the website before a review of the table has been done, and (4) GitHub Actions to extract data from the versioned table and populate website interfaces dynamically. The GitHub repository of *Ragù* project is available online[5], where documentation is available.

Updating the user interface requires an editor to download the working table as .tsv files (including both the metadata and vocabulary tables), and upload them to the project GitHub repository. Committing the uploaded files is labelled as a "data update", and non-expert users can upload the data using a user-friendly drag-and-drop interface. Images are uploaded similarly in the *recipe_photos* folder. These two types of submissions trigger a GitHub action that runs the *script.py* file, which includes methods to extract information from aforementioned files and write a few configuration files. JSON format was chosen since it is lightweight, it is easily manipulated by both Python and JavaScript languages, it is highly readable also by humans, and in the future it can be easily converted into JSON-LD format to add semantics.

The configuration files automatically created/updated by GitHub actions are used to populate the website thanks to a set of JavaScript scripts. These files are:

- general.json contains statistics to be used in the homepage - e.g. the number of recipe books, recipes, and ingredients - and paths to other files (recipe books and filters).
- Other files contain recipes grouped by letter (alphabet.json), type of course (categories.json), ingredients (ingredients.json) and provenance (provenance.json), which are used to populate the advanced search of the website.
- Bespoke files are created to include data necessary to build visualisations, and are divided by type, namely: map.json, matrix.json, network.json and piechart.json, each organised according to the requirements of Charts.js[6], the data visualisation library chosen to create the charts.
- Finally, each cookbook has its own file, in which we find general information such as title, year, origin, and a dictionary including all the recipes, which in turn contain the annotations extracted from the working table.

Such an approach has several advantages, namely: (1) it leverages the power of the hosting platform, hence preventing us from developing sophisticated custom code for ensuring the continuous update; (2) it prevents non-technical people involved in the project to understand complex concepts related to data management and web development, e.g. it does not require them to learn sophisticated software solutions for crowdsourcing (tables are relatively easy to understand), update (drag-and-drop interfaces hide complexities of version-control-systems), and publication (the web publication is completely automated); (3) it is a free-of-charge, potentially sustainable in the mid-term, solution for hosting and versioning code (at the moment), which prevents us from reserving resources for maintenance of a dedicated server; (4) allows one to perform editorial control without requiring yet another interface to learn (only collaborators of the repository can upload the versioned tables that trigger the update of the website).

4. A STORYTELLING PROJECT ON COOKBOOKS

To accommodate requirements of different target audiences, the digital collection is presented on a website[7] that offers three macro-sections, namely: an introductory storytelling journey (*homepage*), a search page based on filters and facets (*recipes*), and an ebook-like browser of digitised cookbooks (*cookbooks*, currently under construction).

The *homepage* offers a digital storytelling journey that exploits the narrative element to showcase the content of the collection and stimulate curiosity. The concept idea leverages the red thread of history and tradition, and uses animation and *scrollytelling*[8] to collate several sections that include charts accompanied by a (sequential) introductory text (see Fig. 2). This narrative uses data visualisation elements to present collected data in an intuitive way and offers insights that can be leveraged in specific searches later - following the approach of generous interfaces and "overview first, details on

[5] https://github.com/raguproject/raguproject.github.io
[6] https://www.chartjs.org/
[7] https://raguproject.github.io/
[8] The term *scrollytelling* is a combination of "storytelling" and "scrolling". It refers to a technique used in web design and digital journalism where narrative content unfolds as the user scrolls down a web page. *Scrollytelling* allows users to explore content at their own pace, while maintaining their attention and deepening their understanding of the topic being presented [11].

demand". In detail, the red string graphically conducts users through five topics that invite them to explore recipes and cookbooks sections, showing significant data patterns that can be appreciated by a broad audience, namely:

1) **Overview**. Users are introduced to the topic (cookbooks), the goal of the project (preservation and dissemination) and are provided with an overview of data, e.g. number of cookbooks, recipes, and ingredients.
2) **Provenance**. The journey continues with a map displaying the provenance of collected cookbooks that stimulates empathy (a user will likely look for recipes from their region) showing where the project originated, and how it evolved engaging with more donors of handwritten notebooks (currently the web application includes only a selection of digitised cookbooks and more have to be included).
3) **Ingredients**. Three data visualisations give the visitor an overview of ingredients: recipes grouped by ingredients show the importance of certain ingredients in the Italian diet, the co-occurrence/correlation of ingredients lets users grasp the constituents of the Italian traditional cuisine, and the analysis of units of measures show them how the experience of the cook has a pivotal role in the Italian intangible heritage.
4) **Dishes**. A set table shows all the types of courses that are collected.
5) **Gender**. A final remark is put on the fact that almost all recipes are written by women.

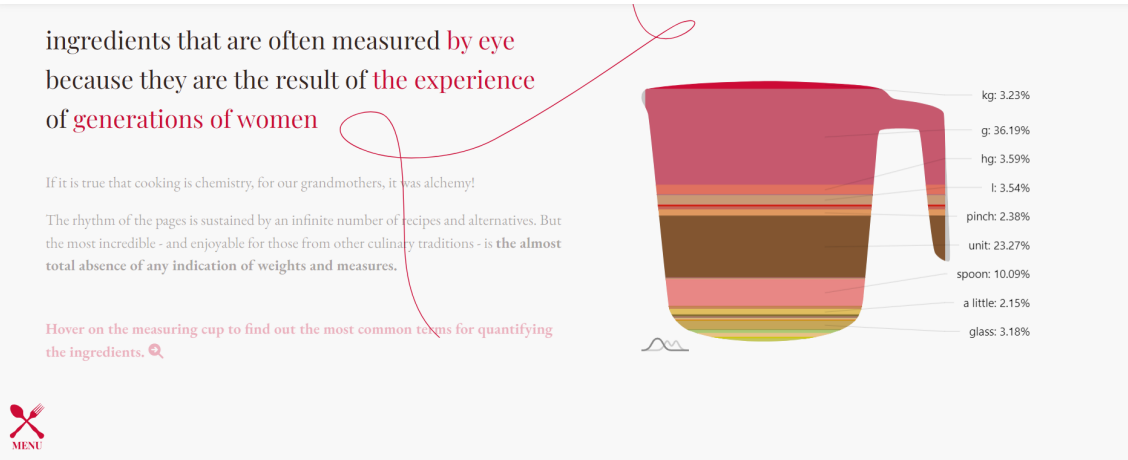


Figure 2. An example of a chart with introductory text

This exploratory journey provides an overall understanding of the collection, informing users of what type of information they could expect later in the exploration of single recipes in the *recipes* section. It is also an engaging way to showcase the collection, the work in progress, and to attract new donors to enrich it continuously.

The second section of the website is the *recipes* where the user can select recipes either from an alphabetical index (like in an encyclopaedia of recipes) or can filter them by category (type of meal, ingredients, city of origin) (see Fig. 3). For each recipe the user can access the digitised picture of the original cookbook and the transcribed metadata (see Fig. 4). If we take as an example the "Pasticcio di maccheroni" recipe we discover two recipes, one by Mrs Dina and the other by Mrs Anna Maria (see Fig. 3). Opening both allows one to appreciate the differences in the making. This section is particularly useful to discover the variants of the names that these women used to call the ingredients, which were the most commonly used and where. Being free in the exploration allows the user to find complex recipes such as "Tortellini" that come in at least 5 different variations, or navigate into poorer recipes, perhaps dating back to war times.

5. CONCLUSION

Despite limited resources, the approach applied to *Ragù*, combining crowdsourcing, good practices in data management, and automated updates through a popular platform such as GitHub, showcases a pragmatic and scalable solution that can be implemented in similar projects. The digital collection's website, structured as an exploratory journey with digital storytelling elements, successfully balances overview and detailed exploration. By leveraging data visualisation techniques, it provides users with insights into the history, provenance, ingredients, dishes, and gender aspects of the cookbooks. The emphasis on an exploratory journey serves not only to enrich understanding but also to invite new contributions, ensuring a dynamic and continually expanding digital collection. Future works will address current work in progress operations, such as the inclusion of new recipes, the finalisation of the ebook-like browser of recipes, and the generalisation of the source code used to build the repository, so that similar projects can leverage out-of-the-box solutions to publish their crowdsourced collections.

Figure 3. Recipes *search interface*

Figure 4. An example of recipe

6. ACKNOWLEDGEMENTS

The *Ragù* project is directed by Dr. Mila Fumini and coordinated by the Digital Humanities Advanced Research Centre (/DH.arc, Unibo): Marilena Daquino (supervision), Giulia Manganelli (web design, web development, copywriting), and Giulia Renda (data management). Data editing: Roberta Balduzzi.
This work was partially funded by Project PE 0000020 CHANGES - CUP J33C22002850006, NRP Mission 4 Component 2 Investment 1.3, funded by the European Union - NextGenerationEU.

REFERENCES

[1] Bonacchi, Chiara, Andrew Bevan, Adi Keinan-Schoonbaert, Daniel Pett, and Jennifer Wexler. 'Participation in Heritage Crowdsourcing'. *Museum Management and Curatorship* 34, no. 2 (2019): 166–82. https://doi.org/10.1080/09647775.2018.1559080.

[2] Bruner, Jerome. 'The Narrative Construction of Reality'. *Critical Inquiry* 18, no. 1 (1991): 1–21.

[3] Daquino, Marilena. 'Linked Open Data Native Cataloguing and Archival Description'. *JLIS* 12, no. 3 (2021): 91–104. https://doi.org/10.4403/jlis.it-12703.

[4] Fernie, Kate, Jillian Griffiths, Mark Stevenson, Paul Clough, Paula Goodale, Mark Hall, Phil Archer, et al. 'PATHS: Personalising Access to Cultural Heritage Spaces'. In *18th International Conference on Virtual Systems and Multimedia,* 469-474, 2012. https://doi.org/10.1109/VSMM.2012.6365960.

[5] Gershon, Nahum, and Ward Page. 'What Storytelling Can Do for Information Visualization'. *Communications of the ACM* 44, no. 8 (2001): 31–37. https://doi.org/10.1145/381641.381653.

[6] Howe, Jeff. 'The Rise of Crowdsourcing'. *Wired*, 2006. https://www.wired.com/2006/06/crowds/.

[7] Kiruthika, Jay, Souheil Khaddaj, Darrel Greenhill, and Jarek Francik. 'User Experience Design in Web Applications'. In IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 642–464, 2016.

[8] Lombardo, Vincenzo, and Rossana Damiano. 'Storytelling on Mobile Devices for Cultural Heritage'. *New Review of Hypermedia and Multimedia* 18, no. 1–2 (2012): 11–35.

[9] Monson, Jane D. *Getting Started with Digital Collections: Scaling to Fit Your Organization.* ALA Editions, 2017.

[10] Navarrete, Trilce. 'Crowdsourcing the Digital Transformation of Heritage'. In *Digital Transformation in the Cultural and Creative Industries*, edited by Marta Massi, Marilena Vecco, and Yi Lin, 1st ed., 99–115. London: Routledge, 2020.

[11] Seyser, Doris, and Michael Zeiller. 'Scrollytelling – An Analysis of Visual Storytelling in Online Journalism'. In *22nd International Conference Information Visualisation (IV)*, 401–6, 2018. https://doi.org/10.1109/iV.2018.00075.

[12] Shneiderman, Ben. 'The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations'. In *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–43, 1996.

[13] Simon, Nina. *The Participatory Museum*. Santa Cruz: Museum 2.0, 2010.

[14] Spennemann, Dirk H. R. 'Datasets for Material Culture Studies: A Protocol for the Systematic Compilation of Items Held in Private Hands'. *Heritage* 6, no. 2 (2023): 1977–85. https://doi.org/10.3390/heritage6020106.

[15] Stevens, Gioia. 'New Metadata Recipes for Old Cookbooks: Creating and Analyzing a Digital Collection Using the HathiTrust Research Center Portal'. *The Code4Lib Journal* 37 (2017). https://journal.code4lib.org/articles/12548.

[16] Whitelaw, Mitchell. 'Generous Interfaces for Digital Cultural Collections'. *Digital Humanities Quarterly* 9, no. 1 (2015).

Serious games e gamification: a che punto sono le istituzioni culturali italiane?

Vincenzo Colaprice
Università degli Studi Bari di "Aldo Moro", Italia - vincenzo.colaprice@uniba.it

ABSTRACT

Nel corso del periodo pandemico le istituzioni culturali hanno sperimentato diverse soluzioni digitali in grado di aumentare la fruizione e l'interazione con il patrimonio culturale. La gamification rappresenta una delle pratiche che sta assumendo una diffusione crescente, affiancata dalla progettazione di videogame collegati a beni ed istituzioni culturali. Un censimento delle soluzioni digitali prodotte dalle cinquecento istituzioni museali italiane più visitate consente di ottenere una fotografia dello stato dell'arte del rapporto tra luoghi del patrimonio culturale e *gamification*. La ricerca scava nel panorama dei *serious games* e della *gamification*, mettendo in luce limiti e opportunità, rivolgendo la propria attenzione al futuro e alla sostenibilità delle soluzioni digitali di gamification nel contesto culturale italiano.

PAROLE CHIAVE

Gamification; videogame; musei; istituzioni culturali.

1. INTRODUZIONE

Secondo il rapporto "I videogiochi in Italia nel 2022", pubblicato dall'Italian Interactive Digital Entertainment Association (IIDEA) nel 2023, tre italiani su dieci sono videogiocatori[1]. Un numero che corrisponde a 14,2 milioni di italiani in età compresa tra i sei e i sessantaquattro anni. Osservando i dati, emerge che l'età media dei videogiocatori italiani è di 29,8 anni e che le donne rappresentano il 42% del totale. Un altro aspetto degno di nota è legato alle tipologie di dispositivi utilizzati, tra i quali primeggia la preferenza per i dispositivi mobili (9,9 milioni di videogiocatori), seguiti da console (6,5 milioni) e PC (5,4 milioni). Infine, è interessante notare la percentuale di videogiocatori in ciascuna fascia di età, risultando particolarmente elevata tra i sei e dieci anni (58%), undici e quattordici anni (71%) e quindi e ventiquattro anni (58%).

Questo complesso di dati è funzionale ad introdurre il tema della relazione sempre più consolidata tra musei, attività culturali e videogame, nonché le ricadute attese in termini di aumento del pubblico e ritorno economico. I processi di *gamification* sono ormai adottati da diverse istituzioni culturali. Nel 2011 Deterding, Dixon, Khaled e Nacke hanno offerto una fortunata definizione di *gamification*: «the use of game design elements in non-game contexts», ovvero l'utilizzo di elementi di progettazione ludica in contesti non ludici [3]. Come ha rilevato Groh, questa definizione si può comprendere a pieno solo considerando le tipologie di applicazioni ludiche impiegate in contesti di varia tipologia con fini educativi, si pensi al mondo dell'informazione, della sanità, dello sport e della cultura [4]. È necessario parlare di *gamification* intendendo il processo di arricchimento dei servizi offerti al pubblico da un qualunque ente, attraverso l'implementazione di caratteristiche che richiamando l'esperienza di gioco, determinano dei risultati comportamentali [5]. Gli effetti della *gamification* vanno nella direzione di un aumento dell'interazione del pubblico e della motivazione nell'apprendimento in diversi contesti, da quelli educativi a quelli culturali. Le soluzioni ludiche prodotte dai processi di *gamification* sono definite *serious games*, ovvero videogiochi che hanno come finalità principale non solo il raggiungimento dell'obiettivo finale del gioco, ma l'apprendimento di determinate nozioni o conoscenze.

Chiariti questi aspetti, si può volgere lo sguardo allo stato dell'arte del dibattito sui processi di *gamification* in Italia. La letteratura scientifica non offre un'analisi complessiva dello scenario nazionale in merito alla presenza della gamification all'interno dei contesti culturali. Al contrario, le pubblicazioni si soffermano maggiormente sull'approfondimento di casi di studio o di buone pratiche. Alcuni studiosi individuano diversi momenti che hanno incentivato l'utilizzo di videogame per la promozione e valorizzazione del patrimonio culturale. Solima [9] individua nella riforma dei musei italiani del 2014 l'origine di un'autonomia gestionale che ha condotto ad un'apertura maggiore dei musei verso l'utilizzo del digitale, superando un approccio che considera le tecnologie digitali come «una mera declinazione in senso dinamico della comunicazione verso l'esterno effettuata dal museo con gli strumenti tradizionali». Pescarin [8] vede nella mostra "Archeovirtual", organizzata a Paestum nel 2017 e dedicata alla relazione tra videogame e archeologia, la svolta che ha spinto i musei a rivolgersi alle aziende del settore videoludico, trasferendo alle imprese gli esisti della ricerca nell'ambito del patrimonio culturale. Lampis [6] evidenzia anche il ruolo avuto dalla decisione UE 2017/864 del 17 maggio 2017 che ha istituito l'Anno europeo del patrimonio culturale, indetto per il 2018 e che invita le istituzioni culturali ad una maggiore

[1] Il rapporto "I videogiochi in Italia nel 2022" è accessibile al seguente link: https://www.iideassociation.com/wp-content/uploads/2024/01/IIDEA_I-videogiochi-in-Italia-nel-2022.pdf.

attenzione nei riguardi delle giovani generazioni, «nella consapevolezza che la loro organizzazione del sapere ha una catalogazione completamente differente da quella del passato millennio».
Questa attenzione si ricollegherebbe all'apertura delle istituzioni pubbliche verso la necessità di un rinnovamento delle pratiche culturali che include anche il digitale tra le metodologie da impiegare, come emerge dal Piano Triennale per la Digitalizzazione e l'Innovazione dei Musei promosso dal MIBAC nel 2018[2]. Seguendo l'approccio della decisione UE, il Piano individua i videogame tra gli strumenti utili a raggiungere gli obiettivi di valorizzazione e marketing culturale, promuovendo la sperimentazione e la ricerca nello sviluppo di *applied games* o *serious games*, ovvero esperienze di gioco che consentono l'apprendimento.
Le limitazioni imposte dalla pandemia di COVID-19 hanno accelerato questo processo, favorendo la creazione di numerose esperienze di gamification, che hanno favorito l'interazione virtuale o a distanza degli utenti con il patrimonio culturale. Uno dei casi più rilevanti a livello internazionale è rappresentato da *Animal Crossing: New Horizons*. Il videogame, che prevede anche la possibilità di giocare in multiplayer online, ha intercettato la congiuntura favorevole realizzata dalla pubblicazione del videogioco nel marzo 2020 e la contemporanea esplosione della pandemia. *Animal Crossing* offre ai videogiocatori la possibilità di esplorare un'isola deserta e modellarla secondo le proprie esigenze, collezionando oggetti e costruendo nuove strutture e ampliando quelle già esistenti sull'isola. Tra gli oggetti collezionabili vi sono anche numerose opere d'arte di musei internazionali, dotate anche di informazioni descrittive dell'opera e relative all'ubicazione nel mondo reale. Il videogame è diventato un vero e proprio best-seller, vendendo circa trentasei milioni di copie in tutto il mondo e riscuotendo un ampio successo di critica [1].
Il caso italiano maggiormente noto proviene dal Museo Archeologico Nazionale di Napoli (MANN), impegnato fin dal 2016 nella sperimentazione di soluzioni legate alla gamification. Il videogame *Father and Son*, pubblicato nel 2017 e prodotto da TuoMuseo, è caratterizzato dalla presenza di alcuni beni del MANN, come la collezione Farnese e manufatti di età egizia e romana, all'interno di una storia a bivi che coinvolge l'utente, offrendo la possibilità di scegliere tra un doppio finale. Il videogame è stato reso disponibile solo per dispositivi mobili[3] e ha raggiunto tre milioni di download nell'anno di rilascio. Il successo riscontrato dal videogame lo ha reso una buona pratica tra le esperienze di gamification adottate dai musei italiani, tenendo in considerazione anche la cura dei dettagli grafici e la presenza di una colonna sonora originale.
Ma quanto sono diffuse queste soluzioni digitali tra le istituzioni culturali? Quanto ha inciso il periodo pandemico sulla diffusione della gamification? Quali istituzioni culturali hanno adottato esperienze di visita "gamificate" o videogame per promuovere le proprie collezioni? Una ricerca quantitativa condotta nell'ambito del progetto di ricerca di dottorato industriale nazionale ha tentato di rispondere a questi interrogativi.

2. METODOLOGIA

L'indagine ha inteso fotografare lo stato dell'arte della gamification nelle istituzioni culturali, realizzando un censimento di videogame ed esperienze di gamification culturale. Le informazioni raccolte hanno consentito di trarre alcune considerazioni a partire dal livello di diffusione di queste pratiche in Italia.
Svolta tra marzo e settembre 2023, la ricerca ha mosso i primi passi con l'individuazione di una base di dati affidabile, stante la volontà di verificare la presenza di videogame o app sviluppate riconducibili alle prime cinquecento istituzioni museali italiane per numero di visitatori. Le cifre relative alle visite sono state ricavate dai microdati prodotti dall'«Indagine sui musei e le istituzioni similari»[4], realizzata da ISTAT nel 2021. I microdati racchiudono le informazioni di oltre quattromilacinquecento istituzioni culturali italiane, classificate in base alle categorie definite da ISTAT: musei, gallerie e/o raccolte; aree o parchi archeologici; monumenti o complessi monumentali; altro.
Redatta la lista delle cinquecento istituzioni museali italiane più visitate, è stata verificata la presenza di soluzioni digitali videoludiche seguendo questi step:

1. Consultazione dei siti internet delle istituzioni culturali e verifica di rimandi a videogame o applicazioni;
2. Consultazione degli shop online, verificando la presenza di videogame legati a beni o esperienze di visita;
3. Consultazione degli store di Google, Apple e Steam per individuare applicazioni legate alle istituzioni culturali censite.

[2] Piano Triennale per la Digitalizzazione e l'Innovazione dei Musei, accessibile al seguente link: http://musei.beniculturali.it/wp-content/uploads/2019/08/Piano-Triennale-per-la-Digitalizzazione-e-l%E2%80%99Innovazione-dei-Musei.pdf..
[3] Sito ufficiale del videogame Father and Son: http://www.fatherandsongame.com/.
[4] Dati accessibili al seguente link: https://www.istat.it/it/archivio/167566.

Questo rilevamento ha consentito di individuare cinquanta istituzioni culturali italiane alle quali sono collegate sessantuno applicazioni ludiche o giochi analogici che integrano l'esperienza di visita. All'interno di un database *flat-file* sono state organizzate le informazioni rilevate per ciascun gioco in base ai seguenti criteri:
a) titolo del gioco; b) istituzione culturale collegata; c) categorie e tipologie delle istituzioni culturali; d) proprietà dell'istituzione culturale (pubblica o privata); e) gestione dell'istituzione culturale (enti pubblici, privati, ecclesiastici o partecipati); f) genere del gioco; g) piattaforma; h) numero di download; i) valutazioni degli utenti (su Google Play, App Store e Steam); l) tipologia (digitale o phygital).

3. RISULTATI DELLA RICERCA

I dati raccolti fotografano un'assoluta predilezione delle istituzioni culturali italiane per applicazioni di natura digitale, soprattutto videogame per dispositivi mobili e console (93,4%), mentre sono ancora del tutto rare le applicazioni phygital, ovvero esperienze di interazione con beni analogici mediate da soluzioni digitali (6,6%). Guardando alle piattaforme sulle quali le applicazioni sono distribuite e per le quali sono sviluppate, emerge una preferenza degli istituti verso la creazione di applicazioni destinate ad essere utilizzate su dispositivi mobili (66,2%), un segnale chiaro della volontà di intercettare il segmento più ampio dei videogiocatori, nonché di fare leva sul dispositivo più facilmente associabile ad un'esperienza di visita. Le esperienze di gaming in realtà virtuale, disponibili sulle piattaforme collegate all'utilizzo di visori VR, rappresentano l'11,8% del totale, seguite dai videogame prodotti per PC (8,8%) e fruibili via web (8,8%). In ultima posizione si classificano i videogame per console (4,4%), un ambito poco esplorato dalle istituzioni culturali italiane. Sono state individuate diverse tipologie di gioco, tra le quali il genere maggioritario risulta essere il videogioco d'avventura (34,7%), spesso coniugato con il genere del rompicapo (13,3%) oppure con un percorso di guida del visitatore (13,3%). Risultano essere secondari generi quali la simulazione (10,7%) o i quiz (9,3%), questi ultimi maggiormente presenti tra le applicazioni web. Il dinamismo della trama e i diversi scenari che connotano il genere dell'avventura si prestano bene allo sviluppo di uno storytelling coinvolgente per l'utente e capace di collocare i beni dell'istituzione culturale all'interno dei passaggi chiave del gioco.
Nel tentativo di valutare il successo di pubblico ottenuto da ciascun titolo, sono state rilevate numerose difficoltà nel raccogliere i dati relativi alle valutazioni attribuite dagli utenti, a causa della scarsità ed enorme difformità di valutazioni presenti sui tre store. Al contrario, è risultato più significativo rilevare il numero di download delle applicazioni. Considerando i primi dieci titoli più giocati, si ricava la seguente classifica:

1) Father and Son (2017), oltre un milione di download (istituzione; MANN; producer: TuoMuseo);
2) A Life in Music (2019), oltre centomila download (istituzione: Teatro Regio di Parma; producer: TuoMuseo);
3) Past For Future (2018), oltre cinquantamila download (istituzione: Museo Archeologico Nazionale di Taranto - MARTA; producer: TuoMuseo);
3) Useeum[5] (2017), oltre cinquantamila download (istituzione: Museo Galileo; producer: Useeum ApS);
5) Cenacolo Vinciano Official App (2019), oltre diecimila download (istituzione: Cenacolo Vinciano; producer: ETT S.p.A.);
5) Father and Son 2 (2022), oltre diecimila download (istituzione: MANN; producer: TuoMuseo);
5) GetCOO Travel (2015), oltre diecimila download (istituzione: Diocesi di Ravenna; producer: GETCOO s.r.l.);
5) Mi Rasna – Io sono etrusco (2018), oltre diecimila download (istituzione: Museo Nazionale Etrusco di Villa Giulia; producer: EGA Ltd.);
5) The Medici Game – Murder at Pitti Palace (2020), oltre diecimila download (istituzione: Palazzo Pitti; producer: Sillabe e TuoMuseo);
5) Museo Teatrale alla Scala (2017), oltre diecimila download (istituzione: Teatro alla Scala; producer: ETT S.p.A.).

Come si può notare, solo due (The Medici Game e Father and Son 2) videogame sono stati rilasciati dopo la diffusione della pandemia di COVID, testimoniando un'attitudine già diffusa tra le istituzioni culturali italiane nel promuovere soluzioni legate alla gamification.
Altri aspetti rilevati riguardano la natura delle istituzioni museali collegate ai giochi censiti e la loro distribuzione geografica. Utilizzando le categorizzazioni assegnate da ISTAT, il 67,2% dei giochi sono prodotti o riconducibili a musei e gallerie, il 14,8% fa riferimento a complessi monumentali, il 9,8% ad aree o parchi archeologici. Approfondendo le tipologie di istituzioni collegate, emerge la prevalenza dell'ambito archeologico (37,8%), rispecchiando la maggiore sensibilità degli enti di questo settore nella creazione di esperienze ludiche relative a beni archeologici. Seguono i musei della scienza (14,8%), musei tematici o specializzati (11,4%), istituzioni legate all'arte (9,9%). Guardando alla

[5] Useeum è un'applicazione che raccoglie audioguide e videogiochi di diverse istituzioni culturali italiane ed europee. Pertanto, il numero di download non si riferisce al contesto esclusivamente italiano.

distribuzione territoriale, il 47,5% dei giochi fa capo ad istituzioni del Nord Italia, il 27,9% al Centro e il 21,5% al Sud. Tuttavia, restringendo lo sguardo alle prime cinque città associate alle istituzioni culturali prese in considerazione, Centro e Sud non appaiono marginali. Milano è la città maggiormente associata ai giochi censiti (21,3%), seguita da Roma (8,2%), Napoli (6,6%), Firenze, Taranto e Torino a pari merito (4,9%). Le città che hanno dato vita alla maggior parte delle applicazioni sono le stesse che guidano la classifica del turismo culturale, ovvero Roma, Firenze, Napoli, Milano e Torino [2], testimoniando una correlazione tra la presenza di queste esperienze e il grado di attrattività turistica delle istituzioni culturali.

Osservando il rapporto tra macroregioni, solo due applicazioni sono riconducibili ad istituzioni culturali settentrionali, mentre cinque provengono da istituzioni del Centro e tre dal Sud. Inoltre, undici delle sessantuno applicazioni censite sono legate ad istituzioni situate in comuni non capoluogo. Anche in questo caso, sono le regioni del Centro e del Sud a dare origine alla maggior parte delle soluzioni sviluppate (sette su dodici).

Infine, destano interesse i dati relativi alla proprietà delle istituzioni culturali: gli enti pubblici sono il 59% del totale, mentre i privati rappresentano il 26,2%, gli enti a partecipazione pubblica sono il 9,8% e quelli ecclesiastici l'1,6%. Sebbene la stragrande maggioranza sia composta da istituzioni culturali pubbliche, non si deve pensare necessariamente ad istituzioni statali. Queste ultime, infatti, rappresentano appena il 26,2%, mentre il 67,2% è di proprietà non statale, ovvero possedute da enti comunali, provinciali, regionali o fondazioni pubbliche.

4. CONCLUSIONI

La presenza di *serious games* e di esperienze di *gamification* appaiono essere una tendenza consolidata e crescente nel contesto delle istituzioni museali. Guardando agli anni di rilascio di videogame e applicazioni si può notare come solo il 37,7% del totale sia stato pubblicato nel triennio 2020-2022, segnato dalla pandemia di COVID-19. La maggior parte dei titoli è stata rilasciata in uno scenario di normalità, con un picco toccato nel 2019. Questi dati contraddicono la percezione che porta a ritenere l'adozione di soluzioni di gamification come dettata dall'inaccessibilità delle istituzioni a causa della pandemia. Al contrario, le sollecitazioni giunte da enti governativi italiani e internazionali, hanno dato impulso ai processi di digitalizzazione delle istituzioni culturali italiane, stimolando lo sviluppo di queste esperienze.

A fronte delle soluzioni digitali ad oggi disponibili e di quelle che verranno, è bene tenere presenti alcuni aspetti già delineati da Pescarin [7] nella definizione delle «dieci regole d'oro» relative allo sviluppo di esperienze interattive e videoludiche nei contesti museali. Tra queste emergono tre temi chiave.

Il primo è relativo alla manutenzione di applicazioni e videogame. Nel momento in cui si scrive (gennaio 2024), prendendo in considerazione le dieci applicazioni più scaricate, emerge il seguente scenario: solo quattro applicazioni risultano aggiornate tra gli ultimi mesi del 2023 e l'inizio del 2024; due applicazioni sono state oggetto di manutenzione per non più di un anno dopo il rilascio e solo una di queste è stata aggiornata per almeno tre anni; tre applicazioni sono state rimosse dai relativi store. L'assenza di aggiornamenti – e dunque di manutenzione – comporta non pochi problemi a fronte della comparsa di dispositivi mobili e sistemi operativi sempre nuovi, causando l'incompatibilità delle applicazioni. Inoltre, la mancanza di aggiornamenti testimonia la conclusione del progetto o l'esaurimento delle risorse dedicate dall'istituzione culturale alla cura dell'applicazione dopo averla distribuita.

Un secondo aspetto riguarda la necessità di evitare la costruzione di «cattedrali nel deserto», ovvero soluzioni digitali scollegate dal percorso di visita o dalle collezioni esposte in un museo. Parte delle applicazioni possiede un approccio volto a promuovere l'istituzione culturale ponendo i beni al centro della trama di gioco, è questo il caso dei videogame prodotti da TuoMuseo. Poche sono le applicazioni che si pongono l'obiettivo di guidare gli utenti nel percorso di visita attraverso esperienze ludiche. In generale, i videogame collegati alle istituzioni culturali appaiono come *stand-alone*, ovvero applicazioni slegate dal contesto culturale che ha promosso il loro sviluppo, rivelando l'attitudine delle istituzioni a considerare il valore esclusivamente strumentale di queste esperienze.

Il terzo elemento riguarda la comunicazione e la valutazione delle applicazioni rilasciate, presupponendo un monitoraggio dei dati di utilizzo e il rilevamento di dati prima e dopo la distribuzione. Al contrario, come è emerso da un'indagine qualitativa attualmente in fase di rifinitura[6], le istituzioni culturali tendono ad investire risorse nello sviluppo di applicazioni che puntano ad aumentare la capacità di attrarre visitatori, sebbene siano poi sprovviste di dispositivi utili a rilevare l'andamento dell'applicazione o non si occupino di rilevare i dati dalle piattaforme di distribuzione. Inoltre, la difficoltà riscontrata nel rintracciare le applicazioni e i videogame sui siti internet delle istituzioni culturali, denota una scarsa presenza di campagne comunicative adeguate che tradiscono il basso grado di rilevanza conferito a queste soluzioni digitali, annoverate tra i tanti strumenti che possono alimentare le strategie di comunicazione attraverso i nuovi media.

[6] I risultati dell'indagine saranno esposti in un articolo di prossima pubblicazione. I dati sono frutto di una ricerca realizzata nel corso del semestre di collaborazione con l'azienda ETT S.p.A., attiva nello sviluppo di soluzioni digitali per il patrimonio culturale.

Concludendo, questa prima analisi dei processi di *gamification* e della diffusione di *serious games* nel contesto culturale italiano trasmette la sensazione che le applicazioni realizzate dai musei siano considerate come progettualità estemporanee rispetto alle attività consolidate di gestione ordinaria. Fatte le dovute eccezioni, nel complesso le iniziative delle istituzioni culturali sembrano orientate verso la volontà di esplorare le ricadute generate dall'implementazione di soluzioni di *gamification* e *serious games.* La scarsa manutenzione destinata alle applicazioni, la poca attenzione rivolta al monitoraggio del numero di download e l'assenza di un'adeguata campagna comunicativa a supporto dei videogiochi costituiscono le manifestazioni più esplicite di questo timido tentativo.

BIBLIOGRAFIA

[1] Cariati, Mario. «Animal Crossing New Horizons senza freni: 35.9 milioni di copie vendute su switch». *Everyeye.it*, 6 maggio 2021. https://www.everyeye.it/notizie/animal-crossing-new-horizons-freni-35-9-milioni-copie-vendute-switch-515796.html.

[2] Cavallo, Lorenzo, Francesca Petrei, e Maria Teresa Santoro, (a cura di). *Il turismo culturale in Italia: analisi territoriale integrata dei dati*. Roma: ISTAT, 2023.

[3] Deterding, Sebastian, Dixon Dan, Khaled Rilla, e Nacke Lennart. «From Game Design Elements to Gamefulness: Defining "Gamification"». In *MindTrek '11: Proceedings of the 15th International Academic MindTrek Conference, 2011*, 9–15, 2011. https://doi.org/10.1145/2181037.2181040.

[4] Groh, Fabian. «Gamification: State of the Art Definition and Utilization». In *Proceedings of the 4th Seminar on Research Trends in Media Informatics,* a cura di Naim Asaj, Bastian Könings, Mark Poguntke, Florian Schaub, Björn Wiedersheim, e Michael Weber, 39–45, 2012.

[5] Hamari, Juho, Joanna Koivisto, e Harri Sarsa. «Does Gamification Work? – A Literature Review of Empirical Studies on Gamification». In *2014 47th Hawaii International Conference on System Sciences*, 3025–34. Waikoloa, Hi: IEEE, 2014.

[6] Lampis, Antonio. «I videogiochi per conoscere arte e cultura». *Economia della cultura* 3 (2018): 269–74. https://www.rivisteweb.it/doi/10.1446/91288.

[7] Pescarin, Sofia. «Esperienze interattive nei musei: dieci regole d'oro». In *Videogames, Ricerca, Patrimonio Culturale*, a cura di Sofia Pescarin, 89–127. Milano: Franco Angeli, 2020.

[8] Pescarin, Sofia. «Videogames, Ricerca e Patrimonio». In *Videogames, Ricerca, Patrimonio Culturale*, a cura di Sofia Pescarin, 11–13. Milano: Franco Angeli, 2020.

[9] Solima, Ludovico. «Il gaming per i musei. L'esperienza del Mann». *Economia della Cultura* 3 (2018): 275–90. https://www.rivisteweb.it/doi/10.1446/91289.

The Tree of Philosophers: design and implementation of a digital resource for the history of academic philosophy

Guido Bonino[1], Nicola Ruschena[2]
[1] University of Turin, Italy – guido.bonino@unito.it
[2] FINO Consortium, Italy – nicola.ruschena@unito.it

ABSTRACT

Our contribution is a presentation of the Tree of Philosophers (ToP), a digital resource for the reconstruction of academic family trees in the history of philosophy, resulting from an on-going collaborative effort of historians of philosophy. ToP's trees represent specific socio-institutional networks of knowledge transmission, as they are made of lines of academic descent that connect philosophers on the basis of institutionalized master-pupil relations. Descent relations are labelled according to specific models varying with historical and institutional contexts, developed in close collaboration with experts in different historical domains. ToP relies on a simple infrastructure whose core is ToP's relational database, which stores philosophers, relations and labels. ToP data have been retrieved from a variety of institutional and administrative sources, integrated by the exam of professional and biographical sources and by selected parts of available genealogical reconstructions. ToP's sources and criteria for data collection allow the resource to include large amounts of philosophers regardless of their notability, thus providing access to a massive extra-canonical collection of non-famous authors. Dealing with 15000 philosophers (mostly unknown) in the first release of ToP presented many challenges concerning FAIRification issues. Such issues have been managed by mapping ToP philosophers on external repositories of virtual identifiers for authority data, integrating them in ToP's database.

KEYWORDS

Linked Open Data; Research infrastructure; Academic descent; Social history of philosophy; Digital resources.

1. INTRODUCTION

The Tree of Philosophers (ToP) is a digital resource for the reconstruction and representation of academic family trees in the field of philosophy. ToP is an ongoing collaborative project, whose network is growing after the first release of the resource at the end of 2023[1]. The use of the term "tree" may involve some suggestions that are not, rigorously speaking, correct. In fact, it may happen that a philosopher has more than one "parent" (multiple supervisors, compresence of different kinds of master-pupil relationships, etc.), so that what one gets is not a simple tree, but a more complex graph, branching in both directions. Furthermore, the data collected do not give rise to a single connected graph, but rather to a number of unconnected ones.

Keeping track of academic kinship and descent has been popular practice in fields like physics and mathematics: Master-pupil relationships, often based on the Doktorvater role, have been recorded in many universities starting from the 16th century, thus providing records that are now converging in online repositories such as Wikidata[2] and in domain-specific resources such as the Mathematics Genealogy Project[3]. More recently, academic family trees appeared for other disciplines: Neurotree[4] reconstructs the academic genealogy of neuroscientists, and its making soon overlapped with the development of a multidisciplinary academic tree by Princeton researchers[5]. Family trees have been provided for philosophy as well (although with little coverage from both a historical and geographical point of view): Princeton's academic tree has a section devoted to (a part of) USA-related philosophy, while famous Australian philosopher David Chalmers hosts the Australasian Philosophy Family Tree[6] on his personal website.

In this landscape, the aim of ToP is to provide an infrastructure available for both public consultation and support to historical research that is focused exclusively on institutional relations, while allowing for comparison among differently characterized contexts.

Data provided by ToP are expected to support historical research especially in its sociological aspects. Indeed, ToP's graphs represent a specific kind of socio-institutional networks in the transmission of academic knowledge, built by connecting

[1] "Tree of Philosophers". https://treeofphilosophers.it/
[2] "Wikidata". https://www.wikidata.org/wiki/Wikidata:Main_Page
[3] "The Mathematics Genealogy Project". https://www.genealogy.math.ndsu.nodak.edu/index.php
[4] "Neurotree". https://neurotree.org/neurotree/
[5] "The Academic Family Tree". https://academictree.org/
[6] "Tree – David Chalmers". https://consc.net/tree/

philosophers (nodes) by a finite set of types of descent relation (arcs). The use of historical network analysis can be of great use in the reconstruction of phenomena affecting different fields or connecting separate traditions [3], [4]. Moreover, the analysis of academic careers [1] and institutional dynamics of power [2] can provide valuable insights for the reconstruction of the interplay between philosophical traditions.

2. DATABASE STRUCTURE

The core of ToP is a relational database consisting of three tables: the Philosophers table, the Edges table and the Labels table. The schema is shown in Figure 1.

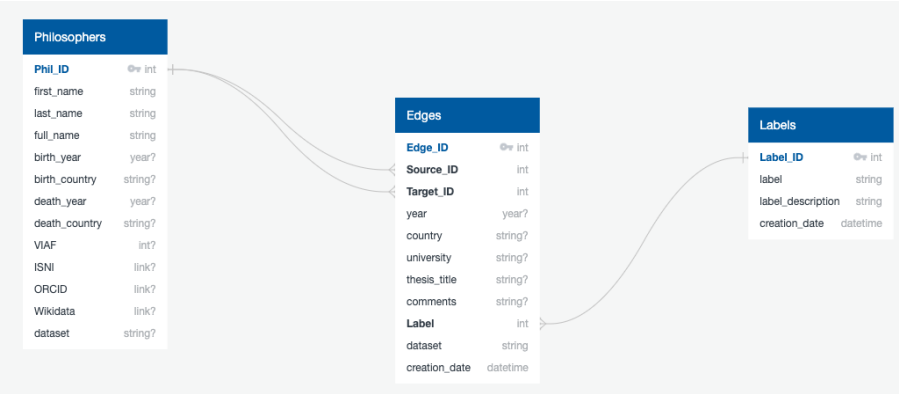


Figure 1

In the "Philosophers" table, rows are indexed by unique numerical identifiers that are internal to ToP's database, which are the primary keys for accessing philosophers' metadata. Each row stores data about one philosopher: name, places and dates of birth and death. As it will be discussed further, external personal identifiers are recorded whenever possible.

The "Edges" table stores data concerning the descent relations between philosophers, which are the core information that forms links in the tree. Since ToP is conceived as an incremental project, the design has been oriented towards simplicity and extensibility: the relation table stores ordered pairs of genealogical "masters" and "pupils" (with foreign keys redirecting to the philosophers' IDs). Noticeably, genealogic lines of descent are unpacked in relations linking two people, a "source" (parent) and a "target" (offspring) of the relation of descent. Each asymmetric relation is indexed by a unique internal identifier and features a year, a country and an academic institution that characterise the specific relation. In those cases in which the relation is connected to the production of a text (e.g. PhD theses for relations of PhD supervision), the title of the text is stored in a specific field, while an additional field is provided to store comments. Comments also store bibliographic references that may be useful to corroborate the descent link, such as snippets of online curricula stating the name of PhD supervisors or archival references. The label assigned to each relation is a foreign key referencing a specific relation type stored in the "Labels" table.

"Labels" table distinguishes between various types of genealogical relationships, such as PhD supervision, graduation thesis supervision, direct tutoring, etc. Each label is assigned a description, which is a textual dossier providing contextual information regarding the historical relation captured by the label and its context of application, provided by the domain experts who identified the specific type of genealogical relationship.

The storage of biographic metadata provides a simple yet effective geographical and chronological individuation of philosophers, while metadata characterizing relations allow researchers to filter their queries by years, countries and academic institutions

These structures allow for the formalization of different historical relations of academic descent, provided that they can be located geographically, chronologically and institutionally. This enables ToP to sustain the addition of a variety of socio-institutional devices of academic filiation characterizing different historical settings[7].

3. INSTITUTIONAL RELATIONS OF DESCENT

ToP focuses on institutional relations, that is to say, on relations that are institutionally recognized, and that can – at least in principle – be documented as such. This means that generic relations of influence, however certain and significant, are not accounted for. The focus is deliberate: ToP is intended as a resource concerning the institutional (which usually means academic) transmission of philosophy. Keeping this kind of transmission distinct from other, more informal, channels allows for possible comparisons, which would be made impossible by mixing things up. It is probably interesting, for

[7] "ToP Labels". https://treeofphilosophers.it/labels

instance, to be able to observe whether the institutional genealogy of philosophy, in specific spatial-temporal circumstances, does or does not approximate the usual historiographic picture of that philosophical context. Moreover, taking generic relations of influence into account would require an appreciable amount of arbitrary decisions for each case. All that would make the Tree of Philosophers the final result of a complex historiographic work, in which the judgments of the editors would play a crucial role. By contrast, we conceive of the tree not as a final result, but as the possible starting point for other researchers; in consideration of both this aim and the collaborative nature of ToP, it is certainly better to keep the role of personal judgments, though well meditated, to a minimum.

ToP started out by considering the most commonly acknowledged historical relation of academic descent, that is the relation between a between PhD candidate and supervisor [8]. Although this kind of relation is common in contemporary academia, many academical environments have been characterized by different institutional relations.

The first part of the project has been devoted to the reconstruction of significant samples from different historical academic contexts characterized by the availability of this kind of relation: 19th- and 20th-century Germany, the United States, Austria, 20th-century Canada, Australia, New Zealand. The volunteer collaboration of a number of historians and PhD candidates in history of philosophy at the university of Turin allowed ToP's scope to grow, encompassing 19th-century French academic training, mid 20th-century Oxford university, 20th-century Italian universities[9]. By relying on specific knowledge of the historical configurations of academic institutions in diverse contexts, collaborators defined historically sound and thoroughly described relations of academic descent, that are used in labelling descent relations in ToP database.

4. PERSONAL IDENTIFIERS

The core data regarding people and relations recorded in ToP are retrieved, whenever possible, from institutional repositories. Such data are integrated with the results of further archival work by ToP researchers browsing archives, on line repositories and domain-specific literature (with an obvious focus on biographical works).

ToP's criteria for the inclusion of people in the Tree of Philosophers are quite broad: anyone who ever granted or received a high-level academic degree *to* or *from* someone who either granted or received a high-level academic degree in philosophy is, in principle, a proper addition to the tree.

People (philosophers) are thus included in the tree regardless of their notability, their career paths, their productivity in the intellectual domain or the reception of their works.

Moreover, a significant part of the tree's domain is populated by what we can naïvely call non-famous philosophers. We cannot provide an esteemed ratio yet (yet!), but common sense is sufficient to assume that in most historical contexts in which academic philosophical training exists, people graduating in philosophy usually come in greater numbers than people becoming notable because of their philosophical work (regardless of training and background).

In some cases, non-famous people have left fewer traces (e.g. because they did not publish philosophical works, having pursued different careers), but what is noticeable in most cases is that those traces that have been left by non-famous people are harder to find and to put together. By relying on resources such as Proquest[10], for example, we can find names of philosophers graduated in the USA in the second half of the 20th century along with the titles of their theses. Nonetheless, such names and titles are seldom sufficient for the attribution of descent relations: apart from cases concerning well-known philosophers, it is very difficult to assess if different pieces of information linked to a name do refer to the same person.

Providing a couple of examples, we retrieved the identities of seven different philosophers whose last name is "Davidson", all of them being trained in the US and active in the 20th century; considering the forty philosophers whose last name is "Johnson" we have been able to distinguish four different "David Johnson" with different and sometimes punctuated middle names, all trained in the US between 1949 and 1978. The attribution to the same people of data retrieved from different sources is often a difficult task simply because we are not sure that they actually refer to the same people. Such attribution thus requires multiple validation steps that are hard, if not impossible, to formalize in a set of instructions. Two

[8] Indeed, this relation is usually the only institutional link that is used by other academic trees. See "Neurotree". https://neurotree.org/neurotree/; "The Academic Family Tree". https://academictree.org/; "The Mathematics Genealogy Project". https://www.genealogy.math.ndsu.nodak.edu/index.php; "Tree – David Chalmers". https://consc.net/tree/

[9] PhD did not exist in the Italian system prior to 1983, while there were no supervisors for French PhDs until 1969. In the Italian case, the supervision of the graduation thesis fits the PhD model well and minor revisions are necessary, such as the addition of one label to distinguish between PhD supervision and graduation supervision. By contrast, the French system did not encompass a one to one direct supervision until rather recently, so that completely different kinds of relations of academic descent had to be considered. For the 19th century, for instance, a reasonable choice is the relation between the lecturer at the École Normale Supérieure in a given year and the students attending the *conférences* in the same year. As for 20th-century Oxford university (at least until the '50s), the most relevant institutional relationship is probably the tutorial in philosophy. See "ToP Labels". https://treeofphilosophers.it/labels. An issue of *DR2 Working Papers* will be devoted to the illustration of some of these case studies.

[10] "ProQuest Dissertations & Theses". https://about.proquest.com/en/dissertations/

major issues that demanded our attention in the development of the Tree of Philosophers are the duplication and the overlapping of the personal identities of philosophers included in the Tree.
We find a partial solution to both problems by relying on virtual identifiers of authority data, which are resources used in archival disciplines and in library institutions. Virtual identifiers are simple numbers or strings of text that are used as labels or indexes, directing to a specific person identified as the author of a number of works. In order to mitigate the problems of duplication and overlapping of identities we made little changes to the database, enlarging the Philosophers Table by adding four additional fields, one for each virtual identifier we chose to rely upon: these are Wikidata[11], ORCID[12], ISNI[13] and VIAF[14].
VIAF (Virtual International Authority File) is an international authority data identifier assigned by aggregating authority data from national library systems: this means that a name that is assigned a VIAF is a name that is recorded as the author of at-least-one work in at-least-one national library catalogue. ISNI is the ISO effort of standardisation of personal identification of contributors to the intellectual production and it works in a similar manner to VIAF, by aggregating authority data from national catalogues along with academic production of article-like works. Both OCLC's VIAF and ISNI aggregates authority data from national systems using samples of published titles and authors' birth (and sometimes death) years. ORCID identifiers are obviously available only for recent times, but their assignment is directly requested by researchers or institutions and they can help in disambiguating nodes in recent branches of the Tree of Philosophers, at the cost of an insignificant increase in the sparsity of the Philosophers Table in the database. Technically, ORCID is a part of ISNI, because ORCID identifiers are included as a region of ISNI identifiers. Nonetheless, we prefer to keep ORCID and ISNI ids as distinct fields. First, because a recent productive philosopher can easily have different ISNI and ORCID ids; secondly, because while ISNI ids are assigned by aggregating data, the assignment of ORCID ids is directly requested by researchers or their institutions, so that the reliability of ORCID is greater than that of ISNI. Finally, Wikidata identifiers are assigned by an automatic system supervised by users of the Wikidata community. Noticeably, Wikidata ids are assigned to somewhat *notable* people so that we cannot expect Wikidata ids to make a huge difference in the disambiguation of non-famous philosophers.
The search for such different identifiers and, whenever available, their attribution to personal records in ToP's database, provide means to improve the way in which ToP works: the inclusion of four types of identifiers makes it possible for us to assess the population of the database at different stages, thus allowing for the evaluation of different strategies of data collection; we can evaluate the coverage of the database in terms of intra- or inter-disciplinary renown of philosophers (e.g. by comparing ISNI, ORCID and Wikidata coverage); furthermore, we are able to approximate a measure of the "Great Unread" that is included in the Tree of Philosophers, by measuring the philosophers that are not recorded in any of the mentioned repositories.
The assignment of virtual identifiers to personal records in ToP's database also improves ToP data in terms of findability, accessibility, interoperability and reusability [5]: if disambiguation allowed by the introduction of virtual identifiers trivially increases findability and accessibility of data, the indexation of philosophers' data with external identifiers dramatically improves interoperability, and consequently reusability. Indeed, reliance on external, widely used identifiers allows for the development of semi-automatic procedures for the inclusion of large quantities of personal records. As stated above, a variety of archival sources provide data about philosophers, and some of these already come in the form of structured data (mostly spreadsheet of archival records). In the best cases, structured data are clean enough to be added almost directly (after some filtering) in ToP's database. Without reliance on external identifiers, we could have done such an addition only if we were sure that no data that we would have added would have overlapped with personal data already present in the database, thus duplicating philosophers. This entails that when two sources of structured data concerning the same context are available, without relying on external identifiers we would have been forced to choose one of the two sources and discard the other. By contrast, reliance on external identifiers and on procedures for their attribution allows for the assignment of identifiers to the structured data to be added, and then for the application of filters in the identifiers' fields.

[11] "Wikidata". https://www.wikidata.org/wiki/Wikidata:Main_Page
[12] "ORCID". https://orcid.org/
[13] "What Is ISNI". https://isni.org/page/what-is-isni
[14] "VIAF". https://viaf.org/

5. INFRASTRUCTURAL INTEGRATION OF ToP – LOD

ToP is going to be one of the pilot projects to be included in the H2IOSC – Humanities and cultural Heritage Italian Open Science Cloud research ecosystem[15] that is being developed by CNR. Under a collaboration agreement between the institute coordinating the Italian node of OPERAS[16] for CNR (Institute for European Intellectual Lexicon and History of Ideas[17]), the Tree of Philosophers project will contribute its results for publication on the platform, participating in the design phases of database models and templates as well as of data acquisition throughout 2024.

The H2IOSC project aims at supporting the transition of participating research infrastructures into a consistent and accessible ecosystem of resources and services providing FAIR data, by creating a federated cloud for historical research in the humanities and for cultural heritage management.

This cloud will offer researchers a single access point to advanced tools, datasets, services, and methodologies provided by ToP and other participating infrastructures, enabling research teams to process, enrich, analyse, and compare research data beyond the boundaries of individual repositories or institutions.

From a practical point of view, the contribution of ToP data to the H2IOSC infrastructure ensures the possibility of their long-term maintenance by an extended ecosystem, rather than relying on the additional, unspecialised (and usually volunteer) work of ToP collaborators.

Moreover, the inclusion in such a large ecosystem of research infrastructures is going to have significant impact on the accessibility and reliability of ToP data for research purposes. Indeed, H2IOSC will provide a stable platform for the publication and management of data, along with means for semi-automatic FAIRness assessment and FAIRification of new data. ToP open data are going to be stored in rdf format, allowing for SPARQL querying[18], and they will be linked with different data created in the context of other research endeavours.

6. ACKNOWLEDGEMENTS

ToP is the result of the project TEPT, funded by Fondazione CRT and carried out by the DR2 research group[19] at the Department of Philosophy and Education science of the University of Turin, in collaboration with the Department of Computer science.

REFERENCES

[1] Bonino, Guido, and Paolo Tripodi. ‘Academic Success in America: Analytic Philosophy and the Decline of Wittgenstein’. *British Journal for the History of Philosophy* 28, no. 2 (2020): 359–92. https://doi.org/10.1080/09608788.2019.1618789.

[2] Katzav, Joel, and Krist Vaesen. ‘The Rise of Logical Empiricist Philosophy of Science and the Fate of Speculative Philosophy of Science’. *HOPOS: The Journal of the International Society for the History of Philosophy of Science* 12, no. 2 (2022): 327–58. https://doi.org/10.1086/721135.

[3] Lalli, Roberto, Riaz Howey, and Dirk Wintergrün. ‘The Socio-Epistemic Networks of General Relativity, 1925–1970’. In *The Renaissance of General Relativity in Context*, edited by Alexander S. Blum, Roberto Lalli, and Jürgen Renn, 15–84. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-50754-1_2.

[4] Petrovich, Eugenio, and Valerio Buonomo. ‘Reconstructing Late Analytic Philosophy. A Quantitative Approach’. *Philosophical Inquiries* 6, no. 1 (2018). https://doi.org/10.4454/philinq.v6i1.184.

[5] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. ‘The FAIR Guiding Principles for Scientific Data Management and Stewardship’. *Scientific Data* 3, no. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[15] “H2IOSC - CNR”. https://h2iosc.ovi.cnr.it

[16] “OPERAS – Open scholarly communication in the European research area for social sciences and humanities ”.. https://operas-eu.org/

[17] “ILIESI Istituto per Il Lessico Intellettuale Europeo e Storia Delle Idee - CNR”. https://www.iliesi.cnr.it/index.php

[18] “SPARQL 1.1 - Overview”. https://www.w3.org/TR/sparql11-overview/

[19] “Distant Reading and Data-driven Research in the History of Philosophy | The Blog of the DR2 Research Group of the University of Turin”. https://dr2blog.hcommons.org/

Thinking Outside the Black Box: Insights from a Digital Exhibition in the Humanities

Sebastian Barzaghi[1], Alice Bordignon[2], Bianca Gualandi[3], Silvio Peroni[4]
[1] Department of Cultural Heritage, University of Bologna, Italy - sebastian.barzaghi2@unibo.it
[2] Department of Classical Philology and Italian Studies, University of Bologna, Italy - alice.bordignon2@unibo.it
[3] Department of Classical Philology and Italian Studies, University of Bologna, Italy - bianca.gualandi4@unibo.it
[4] Department of Classical Philology and Italian Studies, University of Bologna, Italy - silvio.peroni@unibo.it

ABSTRACT[1]

One of the main goals of Open Science is to make research more reproducible. There is no consensus, however, on what exactly "reproducibility" is, as opposed for example to "replicability", and how it applies to different research fields. After a short review of the literature on reproducibility/replicability with a focus on the humanities, we describe how the creation of the digital twin of the temporary exhibition "The Other Renaissance" has been documented throughout, with different methods, but with constant attention to research transparency, openness and accountability. A careful documentation of the study design, data collection and analysis techniques helps reflect and make all possible influencing factors explicit, and is a fundamental tool for reliability and rigour and for opening the "black box" of research.

KEYWORDS

Transparent research; Open Science; Cultural Heritage; Digital twin.

1. INTRODUCTION

In this contribution, we aim to anchor the discussion around open and reproducible research in the Arts and Humanities by presenting as a case study the creation of the digital twin of the temporary exhibition "The Other Renaissance: Ulisse Aldrovandi and the Wonders of the World"[2], currently under development within the PNRR Project CHANGES, and specifically its Spoke 4 – Virtual technologies for museums and art collections [1]. The original exhibition, held in Poggi Palace Museum (Bologna, Italy) between December 2022 and May 2023, consisted of more than 200 objects, mostly belonging to the naturalist Ulisse Aldrovandi and never exhibited before.

The creation of the digital twin – via the acquisition, processing, modelling, export, metadata creation, and upload of the 3D models to a web-based framework – was documented throughout in a structured manner in order to make the entire process transparent and reproducible. Indeed, no reproducibility is possible without transparency, or the careful and complete documentation of all relevant aspects of the study [6: 5].

2. THEORETICAL BACKGROUND AND RELATED WORKS

Goodman and colleagues [6] suggest we should talk about three types of "reproducibility": (i) *methods reproducibility*, i.e. the ability to exactly reproduce a study by using the same raw data and the same methodologies to obtain the same results, (ii) *results reproducibility* – also referred to as *replicability* – i.e. the ability to obtain the same results from an independent study using the same methodologies as the original study, and (iii) *inferential reproducibility*, i.e. "the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study". In explaining how this differs from the two categories previously described, the authors add that scientists might "draw the same conclusions from different sets of studies and data or could draw different conclusions from the same original data, sometimes even if they agree on the analytical results" [6: 4]. The reasons can be *a priori*, such as a different assessment of the probability of the hypothesis being explored, or can be linked to different choices about how to analyse and report data. This third type of reproducibility – which is also the most important according to the authors – might be the most common when talking about research reproducibility in the humanities. Peels and Bouter [12] look at how these concepts can be applied to the humanities, a field that has often been overlooked when talking about reproducible research. They prefer the terms "replicability" and "replication", and again they define three different levels: (i) *reanalysis*, that is Goodman et al.'s *methods reproducibility*, (ii) *direct replication*, where the same study protocol is applied to new data, and (iii) *conceptual replication*, where research data are new and the study protocol is modified [12]. These definitions do not perfectly overlap with those seen before but what is crucial to note here is that the authors find that, while replication can

[1] Authors' contribution according to CRediT (https://credit.niso.org/): Conceptualization (SB, AB, BG, SP); Investigation (SB, AB, BG); Supervision, Validation (SP); Writing – original draft, Writing – review & editing (SB, AB, BG).
[2] https://site.unibo.it/aldrovandi500/en/mostra-l-altro-rinascimento

take various forms across the humanities, it is not fundamentally different from replication in the biomedical, natural, and social sciences and can be achieved by pre-registering the studies, and documenting and sharing methodologies and data [12]. Thanks to specific funding from the research funder NWO[3], a group of Dutch researchers conducted replication studies in a number of disciplines, including the humanities, and recently published a set of recommendations and lessons learned [3]. They found that in all cases replication studies help corroborate the findings of the original studies (e.g., extending the number of sources or using a more state-of-the-art approach) and can provide a more thorough understanding of the relevant research field and the available methodological choices [3: 6-7]. They also note, however, that "even experienced, highly conscientious researchers often find it difficult to document their protocols in enough detail to support direct replication" [3: 8].

Those opposing the application of the reproducibility or replicability categories across all research areas cite the fact that, in several humanities domains, researchers may lack control over the experimental conditions of the original study, or have different viewpoints that produce different data and interpretations [9: 12-13, 13: 7]. Carefully documenting the original study design, data collection and analysis, and reflecting on all possible influencing factors is fundamental for reliability and rigour but does not automatically ensure replicability [13: 10]. Indeed, according to this view, to require replicability of all epistemic cultures is harmful and imposes "universal policies that fail to account for local (epistemic) differences", ultimately denying authority – and related rewards – to researchers in the humanities. On the other hand, the "umbrella of Open Science" is wide enough and its "accountability toolbox" is big enough to develop plural methods for assessing the quality of diverse research practices [13: 12].

Leaving this discussion aside for now, in an increasingly open research environment, well-defined practices are essential for ensuring transparency, reliability, and equitable access to research outcomes. In the case of the digital twin of "The Other Renaissance" exhibition, we looked for some operational indications on how to achieve this goal in the literature produced in research fields relevant to the project. Wilson et al. [16] outline a set of recommendations for scientific computing, applicable across different disciplines and at varying levels of computational expertise. Regarding data management practices, their suggestions focus on the importance of incremental documentation and data cleaning. In particular, they advocate for continuous retention of raw data, robust backup strategies, data manipulation for improving machine and human readability and facilitating analysis, meticulous recording of the steps used to process data, using multiple tables in a way that each record in one table is interlinked with its respective representation in another table via a unique and persistent identifier, and using repositories that issue DOIs to the various data artefacts used and produced for easy access and citation. In the archaeological context, Karoune and Plomp [8] identify three distinct levels of workflow to make research activities reproducible, depending on the computational skills required to carry out such activities. Public access to research materials and methods is facilitated by the first level, which consists of transparent recording through documentation, requiring only the creation and maintenance of a written record of each analysis step, done in a format that allows other peers to read, comprehend, and replicate the work done, while requiring the least amount of computational expertise. Outputs at this level of workflow usually include documents describing the methods and processes, raw data files, and analysis output files. Ensuring version control through a shared file naming system and/or software with history tracking is also a common characteristic of transparent recording since it facilitates the documentation of the process as a whole.

3. MAKING THE DIGITISATION PROCESS MORE TRANSPARENT

To ensure a solid basis for transparency and replicability, our approach closely followed the aforementioned sets of best practices, in line with the indications listed in the Data Management Plan of the project [7]. The digitisation workflow involved creating two datasets as Google Sheet files shared between the team members: one (Object Table, or OT) for storing catalogue descriptions of the physical objects in the collection, the other (Process Table, or PT) for storing data about the digitisation process. After defining the structure of the tables, the variables represented by their headings, and the expected representation for each value, the data were populated in parallel by the team members. On the one hand, the OT was populated with data gleaned from official museum records and preliminary notes related to the exhibition objects, and thus was structured around a cataloguing description of each object (e.g. "title", "author", and so on). Where possible, controlled data values (e.g. people names, terms used for object types, etc.) were aligned with existing vocabularies (such as WikiData[4]) and authority lists (like VIAF[5] and ULAN[6]). On the other hand, the PT was populated with data inserted by

[3] https://www.nwo.nl/en/researchprogrammes/replication-studies
[4] https://www.wikidata.org/
[5] https://viaf.org/
[6] https://www.getty.edu/research/tools/vocabularies/ulan/

the researchers during the acquisition of the objects and the creation of their 3D models and, thus, was structured around the steps involved in the overall digitisation process and their relevant attributes. Overall, the steps include an initial *acquisition activity* for capturing analogue materials and realising their preliminary digital representations, and a series of subsequent activities (*processing*, *modelling*, *export*, *metadata creation*, and *upload*) which involved the use of tools for refining and publishing the 3D models as usable, fully described scientific objects. In turn, these activities were represented as a set of information that included: the organisation responsible for the activity, the people responsible for actually carrying out the activity, the technique and/or tools used to perform the activity, and the timespan in which the activity was carried out. This preliminary work resulted in the creation of a record of the entire digitisation process. Google Sheets and Microsoft Excel in the Microsoft Office 365 platform were strong facilitators for data retention, backup and versioning[7]. Moreover, shared formatting practices on elements such as dates and names were essential for preparing the data for the subsequent phases of the project. At the end of this stage, each object had its metadata, related digitisation phases with their features, and unique identifiers that allowed the two datasets to be linked to each other.

As information was added to both datasets, more work went into getting them ready to be published as machine-readable representations of the entire physical collection, its digital counterpart, and the procedure that, from the former, produced the latter. The Resource Description Framework (RDF)[8] was selected as a formal data representation for enabling transparent data publishing. However, in order to transform the current data into RDF statements, the table structures first had to be mapped to data models that could express and deepen the semantics of the data about cultural heritage and digitisation activities. We chose to reuse the CIDOC Conceptual Reference Model (CIDOC CRM)[9] [4] to represent the data detailing the physical and contextual attributes of the collection objects, and its extension CRM Digital (CRMdig)[10] [5] to depict the stages of the digitisation workflow. The Simplified Agile Methodology for Ontology Development (SAMOD) [14], a methodology to quickly create semantic models that are supported by rich documentation and test cases, was used to draw the needed conceptual constructs from CIDOC CRM and CRMdig and pack them into two data models.

4. MAKING INTERPRETATION MORE TRANSPARENT

One of the main goals of cultural heritage digitisation is the selection of specific elements of reality to store digitally. The selection process involves a deliberate human choice about the physical, geometric, chromatic, mechanical, and stylistic characteristics of the objects to digitise. These aspects are recorded inside a "grid of information", such as vectors, images, 3D models, databases, and tables, among others [2: 127]. According to this logic, a digital technology survey is expected to approximate reality based on some predetermined features selected at the outset of the survey project. The quantity and quality of the data obtained during the survey significantly impact how accurate the digitisation will be. In this context, a *digital replica* is defined as an approximate, aesthetically convincing copy of a cultural site or artefact [2: 127]. In our case study, the main aim was to obtain the digital version of the exhibition's experience, starting from the creation of its digital twin[11], linking to the digital assets of the various objects (3D and multimedia) in the collections, enriched by metadata, catalogued and accessible online using different devices [1: 2].

Our approach for creating the digital twin of Aldrovandi's exhibition included in the first place the implementation of various setups and instruments to create morphologically precise models with highly detailed textures. Photogrammetry and structured light scanner (SLS) acquisition techniques have been used to obtain the digital representation of each item. The choice of these methodologies has been influenced by contextual factors (such as limited time and available space), materials, and the objects' size. We provided documentation about the challenges faced and related solutions adopted in the acquisition and processing phase. The documentation of the risks (e.g. acquisition of non-Lambertian materials, limited object's mobility, etc.) and the solutions adopted (e.g. cross polarisation techniques, specific setup schemas, etc.) permits others to retrace and repeat, at least in theory, the actions involved in a certain research effort, producing new data [15: 2]. Concerning scanner acquisitions, we defined some common limits regarding texture final resolution, and we decided on a specific range for geometry complexity. During the entire process, open technologies and software were employed to maximise the workflow's re-adoption for the creation of a virtual exhibition in different settings. However, for some

[7] Since transparent recording does not involve any computational code, proprietary software like Google Sheets is acceptable as long as it includes features like versioning and exporting outputs to open formats (e.g., .txt, .rtf, .pdf) [7].
[8] https://www.w3.org/TR/rdf11-concepts/
[9] http://www.cidoc-crm.org/cidoc-crm/
[10] https://www.cidoc-crm.org/crmdig/; https://projects.ics.forth.gr/isl/CRMext/CRMdig_v3.2.2.rdfs
[11] Since cultural heritage may be intangible or temporary, Niccolucci et al. [11] suggest separating the data exchange dimension from the representation dimension for digital twins. This reconceptualisation rethinks data flows and bi-directionality as possible and as not mandatory requirements for digital twins of cultural heritage artefacts or landscapes, opening the possibility for accurate digital models (i.e. digital replicas) to evolve dynamically into a fully developed digital twin.

specific tasks (e.g. raw data elaboration), proprietary software was required since open-source software fails to produce satisfactory results.
Documenting processing decisions made for extra transparency should be a part of the scientific workflow and cultural heritage preservation. This can be done, as proposed by Moore et al. [10], by extracting a processing report from the photogrammetry software. Metashape[12] and 3DF Zephyr[13], the main software used for the photogrammetric processing phase, provide this option. The function has not been developed yet for the open-source alternative Meshroom[14], whose implementation in this project is under test. However, software for processing photogrammetric data is considered more open and transparent compared to software used for scanned data elaboration. Scanned data were elaborated using different versions of Artec Studio[15], which has proven to be a "black box" for those who do not own the software and the licence required to use it, allowing raw data export only in proprietary formats and without providing any processing report. Regarding modelling interventions, to guarantee transparency concerning the manipulation of the source data we provided different derivative versions for each 3D model. Level 0 represents the rough result obtained by the acquisition software, while level 1 includes the final high-definition model, where geometry issues have been fixed and lacking parts have been reconstructed. The comparison between these versions enables one to identify which parts were modelled and which parts belong to level 0. Level 2 instead includes the optimised model for web publication. Finally, we used as many standard and interoperable formats as possible for the generated data to facilitate their reuse on different platforms. Specifically, we used glTF, glb, obj, and mtl for 3D models; tiff, jpg, raw, and png for images; mp4 and mov for videos; and mp3 for audios.

5. DISCUSSION AND CONCLUSIONS

We have described how the digitisation process of the exhibition "The Other Renaissance" has been documented throughout, with different methods, but with constant attention to research transparency, openness and accountability. Since any reality-capture or source-based model is affected by the lens of interpretation (of a human or software), tracking steps for the creation of a 3D model is essential to give transparency to these interpretations, facilitating the repeatability of the creation process [10]. Furthermore, data relating to the digitisation process can oftentimes be captured only once, while the process is ongoing, and it is therefore crucial to retain as much information as possible, structure it appropriately and make it available in an open and machine-readable format to provide a record of the entire physical collection, its digital counterpart, and the procedure that, from the former, produced the latter.
A particularly interesting aspect is the temporary nature of "The Other Renaissance" exhibition. At the time of writing, the exhibition concluded more than 6 months ago, objects on loan have been long returned, and the rooms where the exhibition took place have changed use. The same methodologies cannot be applied to the same data – Goodman et al.'s *methods reproducibility* or Peels and Bouter's *reanalysis* – because the physical collection does not exist in its original form anymore. What is possible, however, is that the methodologies described are applied to new data (different cultural heritage objects, exhibitions, etc.). Additionally, the careful documentation of the research process makes it possible for others to judge the relationship between the digital twin and the physical collection, a piece of information that is crucial for scientific scrutiny but that would otherwise have been irremediably lost on the day the temporary exhibition closed.
Documenting the project workflow in this manner is not simple: it requires careful planning, specific competencies, and it is extremely time-consuming. These efforts must be rewarded in the academic setting, if a culture of accountability, data curation and open, reproducible research is to become the norm. Initiatives like CoARA[16] are indeed nudging the scientific community in this direction, but while some practices – such as the publication of research data "as open as possible" and according to FAIR principles – are garnering increasing attention, the focus must be kept on methodologies, too, and on the need of carefully documenting each step of a research project. Further, as noted by Peels and Bouter [12], guidelines on how to report study protocols, methodologies and procedures are needed, and this is perhaps especially true in the humanities. The establishment of principles, like FAIR, and discipline-specific recommendations on how to manage and document research data in a transparent and traceable manner is a great first step in this direction. FAIR principles, supposedly discipline-agnostic, are being discussed and adapted to the different research cultures, Data Management Plans are becoming increasingly common, and templates and online tools are being produced to help researchers fill them out in a structured and machine-actionable manner. There is still more to be done, and more explicit attention needs to be devoted to research methodologies and how to document them in sufficient detail.

[12] https://www.agisoft.com/
[13] https://www.3dflow.net/it/
[14] https://alicevision.org/
[15] https://www.artec3d.com/it/3d-software/artec-studio
[16] https://coara.eu/

We recognise that the debate around the definition of reproducibility and replicability, and whether these terms should be applied to research as a whole, across all disciplines, is not settled [3, 6, 9, 12, 13]. However, there seems to be an agreement on the fact that research can be reproducible in varying degrees, from an "ideal" computational reproducibility all the way to fields where multiple interpretations of a certain phenomenon coexist. Replication here may help "filter out faulty reasoning or misguided interpretations, draw attention to unnoticed crucial differences in study methods" [12] but it is not always possible to ascertain which interpretation is correct. Circling back to the definition of *inferential reproducibility* [6] and the critique of the concept of replicability in a humanities context [9, 13], the researchers' different viewpoints, theoretical background or previous assessments always have a bearing on how the study is conducted and how the results are interpreted. A careful documentation of the study design, data collection, and analysis techniques help reflect and make explicit all possible influencing factors, and is a fundamental tool for reliability and rigour and for opening the "black box" of research.

6. ACKNOWLEDGEMENTS

This work has been partially funded by Project PE 0000020 CHANGES - CUP B53C22003780006, NRP Mission 4 Component 2 Investment 1.3, Funded by the European Union - NextGenerationEU.

REFERENCES

[1] Balzani, Roberto, Sebastian Barzaghi, Gabriele Bitelli, Federica Bonifazi, Alice Bordignon, Luca Cipriani, Simona Colitti, et al. '«Saving Temporary Exhibitions in Virtual Environments: The Digital Renaissance of Ulisse Aldrovandi – Acquisition and Digitisation of Cultural Heritage Objects»'. *Digital Applications in Archaeology and Cultural Heritage* 32 (2024). https://doi.org/10.1016/j.daach.2023.e00309.

[2] Demetrescu, Emanuel, Enzo D'Annibale, Daniele Ferdani, and Bruno Fanini. '«Digital Replica of Cultural Landscapes: An Experimental Reality-Based Workflow to Create Realistic, Interactive Open World Experiences»'. *Journal of Cultural Heritage* 41 (2020): 125–141. https://doi.org/10.1016/j.culher.2019.07.018.

[3] Derksen, Maarten, Stephanie Meirmans, Jonna Brenninkmeijer, Jeannette Pols, Annemarijn de Boer, Hans Van Eyghen, and Gayet Surya, et al. '«Replication Studies in the Netherlands: Lessons Learned and Recommendations for Funders, Publishers and Editors, and Universities»', 2024. https://doi.org/10.31219/osf.io/bj8xz

[4] Doerr, Martin, Christian-Emil Ore, and Stephen Stead. '«The CIDOC Conceptual Reference Model - A New Standard for Knowledge Sharing»'. *Proceedings of the Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modelling, Auckland, New Zealand* 83 (2007): 51–56. https://doi.org/10.13140/2.1.1420.6400.

[5] Doerr, Martin, and Maria Theodoridou. 'CRMdig: A Generic Digital Provenance Model for Scientific Observation'. In *Workshop on the Theory and Practice of Provenance*, 2011. https://api.semanticscholar.org/CorpusID:17819849.

[6] Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis. 'What Does Research Reproducibility Mean?' 8, no. 341 (2016).

[7] Gualandi, Bianca, and Silvio Peroni. 'Data Management Plan: Second Version', 2024. https://doi.org/10.5281/ZENODO.10727879.

[8] Karoune, Emma, and Esther Plomp. 'Removing Barriers to Reproducible Research in Archaeology', 2022. https://doi.org/10.5281/ZENODO.7320029.

[9] Leonelli, Sabina. 'Rethinking Reproducibility as a Criterion for Research Quality'. In Research in the History of Economic Thought and Methodology. Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise, 36B:129–46. Emerald Publishing Limited, 2018. https://doi.org/10.1108/S0743-41542018000036B009.

[10] Moore, Jennifer, Adam Rountrey, and Hannah Scates Kettler. *3D Data Creation to Curation: Community Standards for 3D Data Preservation*. Chicago, Illinois: ACRL, 2022.

[11] Niccolucci, Franco, Béatrice Markhoff, Maria Theodoridou, Achille Felicetti, and Sorin Hermon. 'The Heritage Digital Twin: A Bicycle Made for Two. The Integration of Digital Methodologies into Cultural Heritage Research'. *Open Research Europe* 3, no. 64 (2023). https://doi.org/10.12688/openreseurope.15496.1.

[12] Peels, Rik, and Lex Bouter. 'The Possibility and Desirability of Replication in the Humanities'. *Palgrave Communications* 4 (2018). https://doi.org/10.1057/s41599-018-0149-x.

[13] Penders, Holbrook, and De Rijcke. 'Rinse and Repeat: Understanding the Value of Replication across Different Ways of Knowing'. *Publications* 7, no. 3 (2019): 52. https://doi.org/10.3390/publications7030052.

[14] Peroni, Silvio. 'A Simplified Agile Methodology for Ontology Development'. In *OWL: Experiences and Directions – Reasoner Evaluation*, edited by Mauro Dragoni, María Poveda-Villalón, and Ernesto Jimenez-Ruiz, 55–69. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-54627-8_5.

[15] Rahal, Rima-Maria, Hanjo Hamann, Hilmar Brohmer, and Florian Pethig. 'Sharing the Recipe: Reproducibility and Replicability in Research Across Disciplines'. In *Research Ideas and Outcomes*, Vol. 8, 2022. https://doi.org/10.3897/rio.8.e89980.

[16] Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 'Good Enough Practices in Scientific Computing'. In *PLOS Computational Biology 13*, Vol. 6, 2017. https://doi.org/10.1371/journal.pcbi.1005510.

EDIZIONI SCIENTIFICHE DIGITALI

Il progetto Corr<si>Ca: edizione digitale della corrispondenza Canioni

Anna Giaufret[1], Beatrice Dal Bo[2], Elena Margherita Vercelli[3], Laura Bonanno[4]
[1] Dipartimento di Lingue e Culture Moderne, Dottorato in Digital Humanities, Università di Genova, Italia - anna.giaufret@unige.it
[2] CLESTHIA (EA 7345), Sorbonne Nouvelle, France - beatrice.dal-bo@sorbonne-nouvelle.fr
[3] Dipartimento di Lingue e Culture Moderne, Dottorato in Digital Humanities, Università di Genova, Italia in co-tutela con CY Cergy Paris Université, France - elena.margherita.vercelli@edu.unige.it
[4] Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Dottorato in Digital Humanities, Università di Torino, Italia in co-tutela con l'Université Jean Moulin Lyon 3, France - laura.bonanno@unito.it

ABSTRACT[1]

Il progetto Corr<si>Ca, portato avanti da docenti e dottorande/i del dottorato in Digital Humanities dell'Università di Genova, consiste nella digitalizzazione di una corrispondenza di circa 300 lettere, quella della famiglia Canioni, originaria di Olmi-Cappella, un villaggio dell'entroterra nel nord dell'isola. La corrispondenza si sviluppa nell'arco temporale 1882-1918. Gli scriventi, uomini e donne, presentano diversi gradi di alfabetizzazione nelle due lingue del carteggio, l'italiano e il francese. Il presente contributo presenterà il progetto, non ancora concluso, e i suoi obiettivi, concentrandosi sulla trascrizione delle lettere e sul protocollo XML/TEI utilizzato.

PAROLE CHIAVE

Edizione diplomatica digitale; XML/TEI; corrispondenza; scriventi semicolti; Corsica.

1. INTRODUZIONE: IL CARTEGGIO CANIONI

Il carteggio della famiglia Canioni, l'accesso al quale ci è stato permesso dai discendenti della famiglia stessa, rappresenta una corrispondenza familiare a cavallo tra Ottocento e Novecento, più precisamente tra il 1882 e il 1918. Seppure la corrispondenza si componga di un numero più elevato di lettere, il gruppo di ricerca ha deciso di stabilire una frontiera cronologica – almeno per la prima fase del lavoro – che coincide con un momento chiave della storia europea: la fine della Prima guerra mondiale.

La corrispondenza oggetto di questo studio si compone quindi di 270 lettere, scambiate dai membri e dall'entourage della famiglia Canioni, il cui nucleo centrale si trova in un villaggio della *Haute Corse*, Olmi-Cappella, situato a circa 800 metri di altitudine nella valle del Ghjunsani. Il principale destinatario delle lettere, il figlio maggiore (che è anche colui che ne ha conservato il maggior numero), si trova invece sul continente, vicino a Marsiglia. Il carteggio riflette infatti la lingua parlata sulle due sponde del Mediterraneo da tre generazioni di Canioni, uomini, ma anche donne, appartenenti alla categoria degli scriventi semicolti, con diversi gradi di alfabetizzazione. Inoltre, le lettere sono scritte nel momento in cui la popolazione còrsa passa dall'uso dell'italiano a quello del francese come lingua della scrittura: questo passaggio è visibile nel carteggio.

L'oggetto di studio presenta dunque un interesse linguistico di rilievo, perché ci permette di accedere a dati importanti che riguardano le competenze linguistiche scritte dei semicolti e delle donne al tempo della corrispondenza e la scrittura in zona di diglossia o triglossia. Inoltre, il carteggio costituisce un'importante testimonianza storica: le lettere ci forniscono infatti preziose informazioni sulla vita quotidiana, sulla cultura materiale, sull'economia e la politica locale e su molto altro. Infine, 8 lettere sono inviate dal fronte durante il periodo bellico dal nipote Canioni, il giovane Léon, a cui si aggiungono quelle dei suoi cugini Christophe e Xavier (una quindicina circa): questo gruppo di missive fornisce importanti informazioni sullo stato d'animo dei combattenti e sul rapporto dei Còrsi con la politica nazionale [12].

La geolocalizzazione di mittenti e destinatari delle lettere permetterebbe di acquisire informazioni sull'ampiezza della rete dei corrispondenti, che potrebbero a loro volta supportare riflessioni sulla diaspora còrsa e sul contatto con parlanti che dispongono di repertori linguistici diversi.

[1] Le autrici hanno contribuito rispettivamente alle seguenti sezioni: A. Giaufret (1. Introduzione, 2. Il gruppo e il progetto di ricerca, 3. Metodologia e trascrizione, 4. Particolarità e interesse linguistico del corpus), B. Dal Bo (3. Metodologia e trascrizione, 4. Particolarità e interesse linguistico del corpus), E. M. Vercelli (3. Metodologia e trascrizione, 5. Stato di avanzamento e risultati attesi, Bibliografia), L. Bonanno (2. Il gruppo e il progetto di ricerca, 3. Metodologia e trascrizione, 5. Stato di avanzamento e risultati attesi).

Di seguito viene riportata una rappresentazione grafica dell'albero genealogico della famiglia Canioni (vd. Fig. 1. In rosso i principali scriventi).

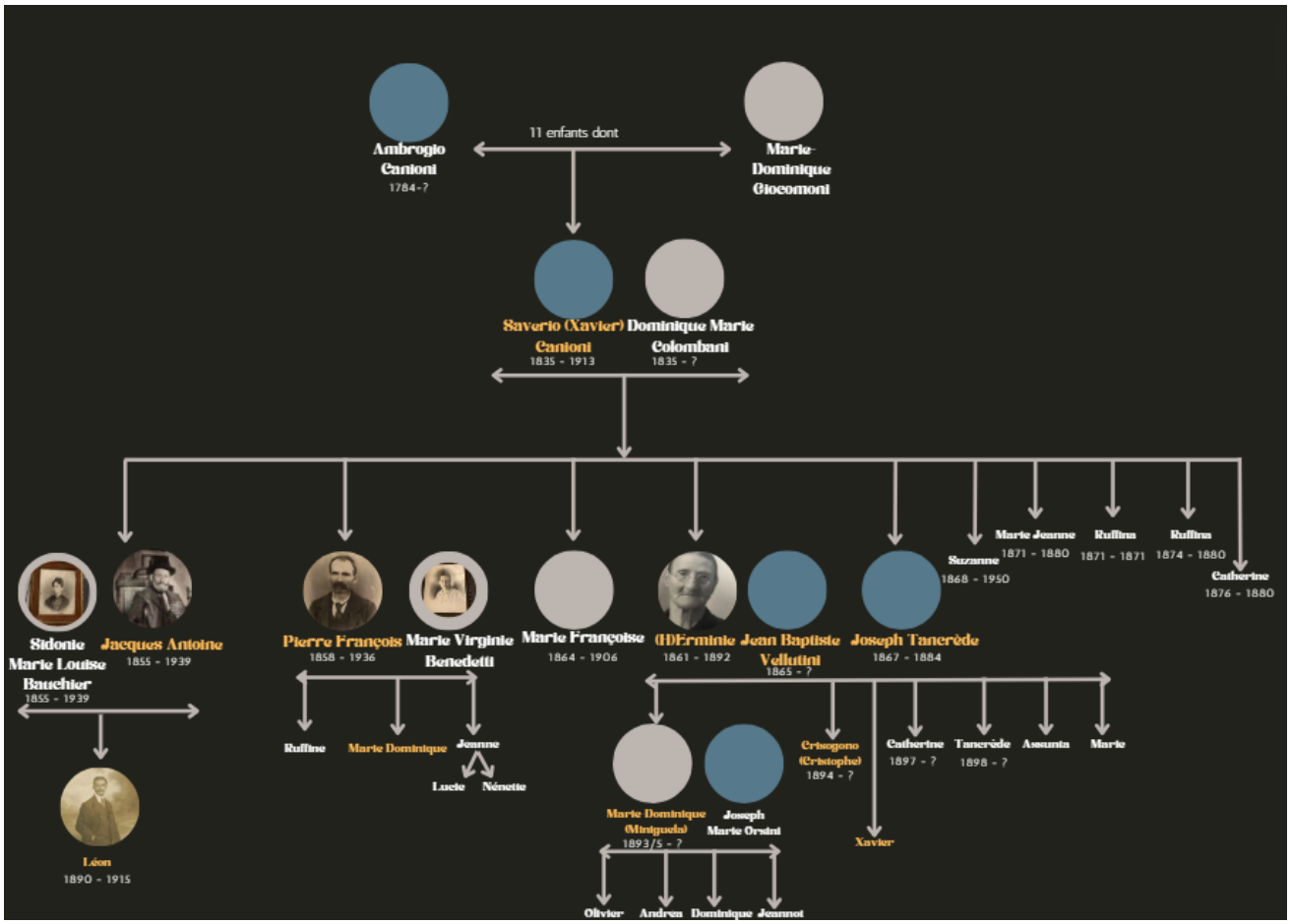


Figura 1. Albero genealogico della famiglia Canioni (in arancione i principali scriventi)

2. IL GRUPPO E IL PROGETTO DI RICERCA

Il progetto ha iniziato a svilupparsi su impulso di Anna Giaufret e Beatrice Dal Bo, prendendo le mosse dal lavoro di ricerca svolto nell'ambito del progetto Corpus 14[2] e nella tesi di quest'ultima, in cui si analizza una corrispondenza di scriventi semicolti durante la Prima guerra mondiale [6, 16]. Sulla base di questa esperienza ed essendo entrate in possesso degli originali del carteggio Canioni, Giaufret e Dal Bo hanno dato inizio al progetto Corr<si>Ca, ovvero Correspondance <numérique> Canioni, e hanno coinvolto le dottorande e i dottorandi in Digital Humanities delle Università di Genova e Torino, ognuna/o delle/dei quali mette al servizio del progetto le proprie competenze. Mentre altri sottogruppi si dedicano all'elaborazione di un'ontologia e alla realizzazione della parte divulgativa del progetto, in una prospettiva di Open Science[3], il gruppo composto da Giaufret, Dal Bo, Bonanno, Vercelli, Mantegazza, Bergaglio[4] si concentra sulla digitalizzazione delle lettere e la trascrizione in XML-TEI.

Il gruppo di ricerca è inoltre in contatto con il *laboratoire M3C* dell'Università di Corsica Pasquale Paoli[5], al fine di rendere fruibile la corrispondenza e i risultati del progetto alla comunità dei Còrsi.

3. METODOLOGIA E TRASCRIZIONE

Non è possibile tracciare nell'ambito di questa presentazione uno stato dell'arte esaustivo dell'edizione digitale (particolarmente fiorente in ambito medievista, di cui è un esempio l'esperienza del codice Pelavicino [18]). In questo

[2] Gruppo di ricerca che ha realizzato la digitalizzazione di un grande corpus di lettere di "Poilus ordinaires" francesi [16, 20] (https://hdl.handle.net/11403/corpus14).

[3] Si veda a questo proposito il contributo Pasciuto *et al.,* «The Corr<si>Ca Project: enhancing and "querying" the Canioni family correspondence» in questo stesso volume.

[4] La trascrizione delle prime 100 lettere è stata realizzata da Giulia Balestrello e Alessia Rebecchi nell'ambito della loro tesi di laurea magistrale.

[5] M3C - Médiathèque Culturelle de la Corse et des Corses: https://m3c.universita.corsica/s/fr/page/home

intervento, ci limiteremo a una rapida panoramica degli studi sulle corrispondenze [9, 10, 14] e sulla digitalizzazione delle stesse secondo lo standard XML-TEI nell'ambito di progetti di ricerca preesistenti, come il progetto Bellini [7].
Secondo Walter et al. [24: 4-5][6], la corrispondenza è un "genere proteiforme, reticolare ed ellittico", caratteristiche che lo rendono particolarmente interessante, ma anche complesso da trattare. Per quanto riguarda la costituzione del corpus del progetto Corr<si>ca, sono state scartate dal carteggio donato dai discendenti della famiglia soltanto le lettere illeggibili (all'interno dell'arco cronologico prescelto). Il carteggio è inoltre accompagnato da materiale fotografico e da altri documenti, quale il libretto militare di Léon Canioni e alcuni taccuini del padre di quest'ultimo, Jacques Antoine, che saranno l'oggetto di una fase successiva del lavoro.
Come osserva Tomasi [23: 131], la "trascrizione delle fonti primarie è una forma di intervento interpretativo" in cui si selezionano e si identificano gli elementi più adatti per rappresentare al meglio la fonte in oggetto. Lo standard XML si rivela particolarmente adeguato a questo compito, in quanto permette ai ricercatori di elaborare documenti *well-formed*, la cui codifica riflette le intenzioni soggettive dell'interprete. Per quanto riguarda gli obiettivi del progetto Corr<si>Ca, la volontà è quella di mettere a disposizione della comunità dei ricercatori e della società in generale un corpus di lettere consultabile per ricerche di tipo linguistico innanzitutto, ma anche storico, filologico, ecc. Per svolgere questo lavoro in modo conforme allo standard di codifica di testi digitali, utilizziamo lo schema di validazione della TEI[7], che permette la validazione dei documenti, che sono dunque non soltanto "ben formati", ma anche "validi" (*valid*), ossia rispondenti agli standard previsti dal consorzio [3]. Partendo da tali obiettivi e rispettando i principi FAIR (*Findability, Accessibility, Interoperability, and Reusability*) [25], la prima fase del lavoro ha permesso di definire i principi alla base della trascrizione e gli elementi con relativi attributi e valori.
Per la trascrizione è stata utilizzato un protocollo di codifica XML-TEI, creato da Dal Bo a partire dalle informazioni accessibili di altri progetti di ricerca che lavorano su tali oggetti di studio e arricchito degli elementi necessari alla codifica delle particolarità linguistiche di questo corpus, al fine di facilitare e uniformare il più possibile il lavoro dei vari collaboratori al progetto. Questi hanno ugualmente accesso alla versione 2.0 del protocollo redatta in francese e aggiornata da Dal Bo e Giaufret nel novembre 2023. Esso si compone di tre sezioni: A) *TEI Header*, B) Organizzazione testuale della lettera, C) Corpo del testo. La prima (A) presenta alcune delle etichette da utilizzare per l'annotazione dei metadati, fra cui alcuni elementi di `<correspDesc>` [19]. La seconda sezione (B) presenta i tag che identificano le varie parti che compongono la lettera, come `<opener>` e `<closer>` per le formule epistolari e `<postscript>` per eventuali *post-scriptum.* Infine, la terza (C) contiene i tag necessari alla riproduzione dei documenti originali, sia per quanto riguarda la struttura delle sezioni del testo (ad esempio, `<p>`, `<pb/>`, `<lb/>`, ecc.), sia per le modifiche apportate dagli scriventi nel processo di scrittura (`<del>`, `<add>`, etc.) con i relativi attributi. In questa sezione sono anche indicate le etichette che possono essere utilizzate per annotare elementi specifici, che sono stati precedentemente identificati come oggetti di analisi linguistiche pertinenti (per esempio, `<unit>` per le unità di misura, `<foreign>` e `<span>` in caso di uso di utilizzo di lingue diverse da quella principale del documento) e per l'analisi di contenuti specifici, quali i nomi di persona e di luogo (`<persName>`, `<placeName>`) per le entità nominate corrispondenti.
È stato inoltre creato un database che viene costantemente aggiornato man mano che le trascrizioni sono completate e caricate nell'archivio. Tale database contiene colonne utili a identificare per ogni lettera elementi quali la lingua principale in cui è scritta, mittente e destinatario, ma anche luogo di partenza, luogo di destinazione e luoghi menzionati. È per questo una fonte importante di dati linguistici e geografici, che saranno utili per rapide ricerche e per la creazione delle cartine geografiche che accompagneranno l'edizione.

Alla luce delle problematiche emerse durante la fase di trascrizione e annotazione del corpus, un'ulteriore versione del protocollo di codifica XML/TEI sarà definita nell'autunno 2024. Le difficoltà riscontrate riguardano principalmente la natura stessa del linguaggio di markup, che, essendo costruito su una struttura ad albero rigidamente gerarchica [3], non sempre permette di trascrivere con la fedeltà che ci si impone in seno al progetto le particolarità testuali di alcune lettere (ad esempio la gestione di più elementi `<closer>` o l'inserimento di questi nel margine del foglio). La versione aggiornata del protocollo sarà quindi applicata all'insieme del corpus nella fase di lavoro successiva, quella della revisione e armonizzazione delle trascrizioni, prevista per l'estate 2025. La versione definitiva del protocollo sarà caricata nell'archivio digitale Zenodo. In questa versione aggiornata è prevista l'integrazione esplicita con l'ontologia di cui sopra, specialmente per quanto riguarda le classi create, che si ispirano direttamente alle etichette utilizzate nella fase di trascrizione. Lo scopo di tale lavoro è di rendere più trasparenti le scelte operate dai due gruppi che lavorano sulla codifica e sull'ontologia e il legame tra i due protocolli.

[6] L'Édition numérique de correspondances https://cahier.hypotheses.org/guide-correspondance
[7] TEI Guidelines: https://tei-c.org/guidelines/

Riprendendo la distinzione di Pierazzo [15] tra edizioni *haute couture,* che prevedono l'uso di un software creato ad hoc per gli interventi a valle dell'edizione, e edizioni *prêt-à-porter,* che implicano, invece, l'adattamento di un software esistente, il progetto Corr<si>Ca si colloca nel secondo gruppo. Infatti, si prevede di utilizzare uno strumento come EVT (*Edition Visualization Technology*) per la visualizzazione e TXM[8] [13] o un altro strumento equivalente per le interrogazioni complesse di tipo linguistico nel corpus. In parallelo, lo sviluppo di un blog Wordpress permetterà di valorizzare conoscenze e di inserire strumenti di accompagnamento (per la geolocalizzazione, la visualizzazione della rete di scriventi, ecc.).
La scelta di EVT (ancora in fase di valutazione) come software per la creazione, navigazione e visualizzazione dell'edizione digitale della corrispondenza sarebbe motivata in primis dalla compatibilità con gli standard XML, HTML e Java, che permettono flessibilità e configurabilità [8, 17] e, in secondo luogo, dalla possibilità di sfruttare funzionalità di geolocalizzazione già presenti in EVT 2, ma che verranno implementate con migliori collegamenti tra testo, risorse LOD e mappe in EVT 3, la cui versione alpha è stata lanciata nel dicembre 2022 e che probabilmente verrà rilasciato nel 2024.
Precisiamo che l'edizione ha escluso per il momento di interessarsi agli aspetti materiali delle lettere (grana della carta, penna, timbri, ecc.), salvo quando questi elementi possono costituire indizi utili al collocamento temporale o geografico della missiva.

Nella prima fase del progetto, ci siamo interrogate sulla granularità della codifica, su un eventuale protocollo di anonimizzazione (per i dati sensibili) e sul tipo di edizione. Abbiamo infine optato per una trascrizione diplomatica delle lettere che rispetti criteri di fedeltà rigorosa all'originale (segmentazione, punteggiatura, uso delle maiuscole, ortografia, ecc.).

4. PARTICOLARITÀ E INTERESSE LINGUISTICO DEL CORPUS

Oltre alle caratteristiche della lingua utilizzata dagli scriventi semicolti in queste missive, la fase di trascrizione e codifica ha messo in evidenza altre particolarità del corpus, frequenti nei corpora di corrispondenze di persone "comuni", quali l'alto numero di scriventi non presenti nelle basi di dati del web semantico (come VIAF[9]) e i nomi di luogo assenti nei database di geolocalizzazione come DBpedia[10] e GeoNames[11], perché poco noti, locali e/o dialettali. Inoltre, poiché non esiste un'edizione critica precedente delle lettere, sono emersi elementi problematici, alcuni dei quali tipici dell'edizione di corrispondenze (elementi non decifrabili o non presenti, numerose grafie da decifrare e/o identificare, presenza di lettere non autografe, ecc.), alcuni più specifici alle corrispondenze di scriventi semicolti [20][12] (usi linguistici lontani dalla norma dell'epoca), altri ancora specifici alla regione di provenienza, poiché si tratta di scriventi bilingue o trilingue, le cui lettere costituiscono una preziosa testimonianza del momento di transizione linguistica dall'italiano al francese, con l'emergenza, talvolta, del substrato còrso. Infatti, la Corsica attraversa tra l'Ottocento e il Novecento una fase di francesizzazione, scandita da momenti più o meno intensi, e che agisce soprattutto sulla popolazione scolastica. Secondo alcuni studi sull'argomento [1, 22], il reale arretramento dell'italiano lingua scritta di fronte all'avanzata del francese avviene solo intorno alla metà dell'Ottocento, mentre la trasmissione intergenerazionale della lingua còrsa si mantiene fino agli anni Cinquanta del Novecento. La situazione linguistica della popolazione può quindi essere definita di diglossia, con una lingua parlata, il còrso, e una lingua scritta che è dapprima l'italiano e poi il francese, con una fase di transizione in cui il repertorio linguistico si compone di tutte e tre, come nel caso di Pierre François. Il nostro corpus contiene 8 lettere scritte interamente in italiano, tutte da Xavier Canioni (nato nel 1835) e due *post scripta*, uno di Pierre François (nato nel 1858) e uno di Xavier. Dall'analisi delle grafie abbiamo ipotizzato che le lettere in francese di Xavier (dal 1888 in poi, con qualche eccezione) siano in realtà dettate al figlio Pierre François, che le traduce dal còrso o dall'italiano, ma che è tuttavia in grado di scrivere anche in quest'ultima lingua.
Gli esempi che seguono mostrano non solo come la lingua usata, sia essa italiano o francese, si discosti dalla norma dello standard, ma anche come si manifesti la presenza di lemmi della lingua còrsa (con *code-mixing*), soprattutto laddove gli argomenti affrontati sono relativi alla cultura materiale. Peraltro, l'uso del còrso non si limita ai nomi degli alimenti (in questo esempio specifico), ma si diffonde anche agli articoli e ai numerali:

[8] TXM: https://txm.gitpages.huma-num.fr/textometrie/index.html
[9] VIAF: https://viaf.org/
[10] DBpedia Fr: https://fr.dbpedia.org/
[11] GeoNames: https://www.geonames.org/
[12] Inoltre, per studi sulla lingua dei semicolti, rimandiamo a [2] per il francese, [5, 11, 21] per l'italiano e [4] per l'italiano in Corsica.

“Nous vous avons expédier un petit colis postal de cinque kilos. contenant una salticca, deux fiadelli una pulpetta, una fetta di mezina e dui furmagli. Voilà le tout, nous avons coupé un petit morceau de pulpetta qui passait le poid”. (02/02/1914, Pierre François; si noti anche l’anno, che si colloca alla fine dell’arco cronologico del corpus).

Ecco, quindi, la prima proposta di codifica adottata dal progetto di ricerca:

```
<p>Nous vous avons expédier<lb/>
un petit colis postal<lb/>
de cinque <unit type="weight" unit="kg"> kilos</unit>. contenant<lb/>
<foreign xml:lang="co">una salticca</foreign>, deux<lb/>
<span xml:lang="co">fiadelli una pulpetta,<lb/>
una fetta di mezina<lb/>
e dui furmagli</span>.<lb/>
Voilà le tout, nous avons<lb/>
coupé un petit morceau<lb/>
de <foreign xml:lang="co"> pulpetta</foreign> qui passait<lb/>
le poid.</p>
```

Altro esempio interessante delle specificità del corpus è quello che riguarda le unità di misura, per le quali gli scriventi continuano a usare quelle tradizionali (in particolare *rubo* e *cantaro*) anche quando il sistema decimale (usato anch’esso) si è diffuso in Francia, e quindi in Corsica:

“Mi dimandi se le uve sono bone. Si, sono bone, e molto a marchato Ecco il prezzo, 30 soldi il rubo. Cioe 6, franchi il cantaro dunque sapia che ci vole, 13 rubi duva per fare, 10020 litri di mosto. Viene dunque a, 3 soldi il litro. Il prezzo e bono”. (30/01/1881, Xavier)

Questa la proposta di codifica della citazione precedente:

```
<p>Mi di mandi sele uve sono bone<lb/>
si sono bone e molto à marchato<lb/>
E cco il prezzo. 30 <unit type="currency">soldi</unit> il <unit
type="weight">rubo</unit><lb/>
cioe. 6, <unit type="currency">franchi</unit> il <unit
type="weight">Cantaro</unit> dunque<lb/>
sapia che ci vole, 13, <unit type="weight">rubi</unit> duvaper<lb/>
fare, 100 20 <unit type="volume" unit="L">li tri</unit> di mosto.</p>
<p>viene dunque a, 3. <unit type="currency">soldi</unit> il <unit type="volume"
unit="L">litro</unit>.<lb/>
il prezzo e bono.</p>
```

Esistono infine esempi di lettere scritte nelle due lingue, ovvero in cui il corpo della lettera è in francese e il *post scriptum* in italiano. L’elemento interessante è che queste due lettere sono entrambe dello stesso scrivente, Pierre François, ma nel primo esempio è lo stesso Pierre François a esserne l’autore (20/09/1887), mentre nel secondo la lettera è firmata da Xavier, sebbene la grafia sia attribuibile allo stesso Pierre François (30/06/1893, sei anni dopo la precedente).
Ecco la proposta di codifica di quest'ultimo esempio (lettera di Xavier, scritta da Pierre François, 30/06/1893):

```
<postscript>
<p><span xml:lang="it">Caro figlio mi vene per<lb/>
iscontro un cabriole a unpr<lb/>
ezzo tutto affatto à mercato<lb/>
per di meglio per 90 <unit type="currency">franchi</unit><lb/>
che avendolo da comprare in<lb/>
fabrigo costarebe almeno 150 <unit type="currency">fra<add
place="above">n</add>chi</unit>. ma non trovandomi<lb/>
```

```
mica instato di denaro per<lb/>
podere fare questa compra<lb/>
mi adirizzo a te si nulla ti<lb/>
pergudigueca et se ti le trovi<lb/>
a la mano fendodi il tuo<lb/>
bougletto <del rend="overstrike">et et</del> e il piu presto<lb/>
sarai pagato frutti et fondo<lb/>
Se tu decili : per la soma di 100 <unit type="currency">fr</unit> fr.<lb/>
rispondi il piu presto possibile</span></p>
</postscript>
```

Dagli esempi precedenti si evincono le difficoltà di codifica a causa di problematiche linguistiche relative alla segmentazione dei lemmi, alla distanza dalla norma generalizzata e alle interferenze continue tra le tre lingue in compresenza. Tuttavia, è proprio da queste difficoltà che potranno nascere le riflessioni più innovative sul piano scientifico. Per esempio, si prevede di utilizzare il tag `<w>` per indicare la segmentazione standard delle parole che si trovano ipo- o ipersegmentate nel testo originale e `<choice>` per segnalare le forme standard. Seguendo le raccomandazioni della TEI, l'obiettivo è di indicare, all'interno di `<choice>`, la forma originale attraverso il tag `<orig>`, e la forma standard utilizzando l'elemento `<reg>` e l'attributo `@type` per precisare il livello linguistico interessato dalla normalizzazione (per esempio, ortografia, morfologia, sintassi).

```
Mi <w>di mandi</w> <w>se</w><w>le</w> uve sono bone<lb/>

Nous vous avons <choice><orig>expédier</orig><reg
type="morphosynt">expédié</reg></choice><lb/>
```

Ulteriori difficoltà di codifica sono presentate dall'organizzazione grafica del testo, le cui sezioni non sono sempre delimitate chiaramente, dall'occupazione da parte di questo di molti spazi a margine, dall'ordine testuale che non corrisponde all'ordine grafico, ecc.

5. STATO DI AVANZAMENTO E RISULTATI ATTESI

Le fasi del progetto già realizzate e previste sono quindi:

1. Elaborazione e firma di un protocollo d'intesa con i discendenti della famiglia Canioni (conclusa)
2. Scansione delle lettere in HD (terminata)
3. Elaborazione del protocollo di codifica (prima versione realizzata)
4. Trascrizione delle lettere e codifica in XML-TEI (60% circa)
5. Revisione delle trascrizioni (da effettuare alla fine della fase precedente e della fase 6.1)
6. Codifica strutturale e semantica
    6.1. Fase preliminare: metadati, testo, unità di misura, nomi di luogo e di persona (in corso)
    6.2. Fase avanzata: inserimento di etichette per standardizzare le forme e annotazione semantica delle entità nominate (attesa di finanziamenti supplementari).

Parallelamente, si sta lavorando alla firma di un accordo con l'Università di Corsica Pasquale Paoli, affinché il progetto possa essere ospitato sul sito del Laboratoire M3C ed essere reso disponibile alla comunità dei ricercatori e degli appassionati in Corsica.

I risultati attesi alla conclusione del progetto saranno quindi un portale web da cui saranno accessibili: la piattaforma di visualizzazione delle lettere, accompagnata dall'edizione diplomatica; il corpus su TXM o strumento simile per effettuare ricerche linguistiche complesse tramite etichette XML-TEI; un blog par la valorizzazione della ricerca che conterrà un apparato critico e divulgativo, compreso quello iconografico e documentale.

Siamo convinti che questo materiale potrà essere utile alla comunità degli storici, dei linguisti, degli etnografi, ecc. per approfondire la conoscenza di una microregione còrsa, il Ghjunsani, dei suoi rapporti con altre regioni dell'isola e del continente e di un periodo storico fondamentale per la storia del repertorio linguistico dei Còrsi.

6. RINGRAZIAMENTI

Ringraziamo i discendenti della famiglia Canioni, che ci hanno generosamente prestato la corrispondenza originale oggetto di questo studio, e Santu Massiani, prezioso custode della storia del Ghjunsani.

BIBLIOGRAFIA

[1] Branca, Marina, e Nicolas Sorba. *«Un siècle d'évolution de la transmission intergénérationnelle du corse»*. Glottopol 38 (2023). https://doi.org/10.4000/glottopol.3179.

[2] Branca-Rosoff, Sonia, e Nathalie Schneider. L'Écriture des citoyens. Une analyse de l'écriture des peu-lettrés pendant la période révolutionnaire. Paris: Klincksieck, 1994.

[3] Ciotti, Fabio, (a cura di). *Digital Humanities. Metodi, strumenti, saperi*. Roma: Carocci, 2023.

[4] Colombani Giaufret, Hélène, e Anna Giaufret. «Il manoscritto dei verbali del comune di Pioggiola (Corsica), 1788-1797: analisi testuale e linguistica». In *Una piccola comunità corsa negli anni della Rivoluzione. Pioggiola attraverso il manoscritto delle delibere 1787-1797*, a cura di Francesca Ferrando, 35–59. Palermo: New Digital Press, 2022.

[5] D'Achille, Paolo. «L'italiano dei semicolti». In *Storia della lingua*, a cura di Luca Serianni e Pietro Trifone, 2:41–79. Torino: Einaudi, 1994.

[6] Dal Bo, Beatrice, Francesca Frontini, e Giancarlo Luxardo. «Annotazione semantica e visualizzazione di un corpus di corrispondenze di guerra». In *Atti del IX Convegno Annuale AIUCD. La Svolta Inevitabile: Sfide e Prospettive per l'Informatica Umanistica*, (a cura di) Cristina Marras, Marco Passarotti, Greta Franzini, e Eleonora Litta. Quaderni di Umanistica Digitale, 2020. https://doi.org/10.6092/UNIBO/AMSACTA/6316.

[7] Del Grosso, Angelo Mario, Erica Capizzi, Salvatore Cristofaro, Maria R. De Luca, Emiliano Giovannetti, Simone Marchi, Graziella Seminara, e Daria Spampinato. «Bellini's Correspondence: a Digital Scholarly Edition for a Multimedia Museum». *Umanistica Digitale* 3, fasc. 7 (2019): 23–47. 10.6092/issn.2532-8816/9162.

[8] Di Pietro, Chiara, e Roberto Rosselli Del Turco. «Between Innovation and Conservation: The Narrow Path of UI Design for the DSE». In *Digital Scholarly Editions as Interfaces*, a cura di Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, e Gerlinde Schneider, 12:133–63. Schriften Des Instituts Für Dokumentologie Und Editorik. Norderstedt: Books on Demand, 2018.

[9] Donato, Maria Pia. «Lettere, corrispondenze, reti epistolari». *Mélanges de l'École française de Rome - Italie et Méditerranée modernes et contemporaines* 132, fasc. 2 (2020): 249–55. https://doi.org/10.4000/mefrim.9995.

[10] Duménieu, Bertrand, Danièle Poublan, Génério Jean-Damien, Francine Filoche, e Patricia Bleton. «Un wiki sémantique pour l'édition scientifique d'une correspondance du XIXe siècle». *Humanités numériques* 6 (2022). https://doi.org/10.4000/revuehn.3203.

[11] Fresu, Rita. «L'italiano dei semicolti». In *Manuale di linguistica italiana*, a cura di Sergio Lubello, 328–50. Berlin, Boston: De Gruyter, 2016.

[12] Géa, Jean-Michel. «Entre identité locale et sentiment national: la posture énonciative de deux soldats corses durant la Première Guerre mondiale». *Études corses* 59 (dicembre 2004): 129–43.

[13] Heiden, Serge, Jean-Philippe Magué, e Bénédicte Pincemin. «TXM: Une plateforme logicielle open-source pour la textométrie – conception et développement». In *JADT 2010: 10th International Conference on the Statistical Analysis of Textual Data*, 1021–32. Roma, 2010.

[14] Martineau, France, e Sandrine Tailleur. «Correspondance familiale acadienne au tournant du XXe siècle: fenêtre sur l'évolution d'un dialecte». In *Congrès Mondial de Linguistique Française*. CMLF, Paris: Institut de Linguistique Française, 2010. https://doi.org/10.1051/cmlf/2010118.

[15] Pierazzo, Elena. «Quale futuro per le edizioni digitali? Dall'haute couture al prêt-à-porter». In *Atti del V Convegno Annuale AIUCD 2016*, a cura di Federico Boschetti. Venezia, 2017. https://doi.org/10.6092/unibo/amsacta/5559.

[16] Praxiling - UMR 5267. «Corpus 14». ORTOLANG (Open Resources and TOols for LANGuage), 2019. https://hdl.handle.net/11403/corpus14/v2.

[17] Rosselli Del Turco, Roberto. «Designing an Advanced Software Tool for Digital Scholarly Editions»: *Textual Cultures* 12, fasc. 2 (2019): 91–111. https://doi.org/10.14434/textual.v12i2.27690.

[18] Salvatori, Enrica, Roberto Rosselli Del Turco, Chiara Alzetta, Chiara Di Pietro, Chiara Mannari, e Alessio Miaschi. «Il Codice Pelavicino tra edizione digitale e Public History». *Umanistica Digitale*, No 1 (2017). https://doi.org/10.6092/ISSN.2532-8816/7232.

[19] Stadler, Peter, Marcel Illetschko, e Sabine Seifert. «Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <corrispDesc>». *Journal of the Text Encoding Initiative*, fasc. 9 (2016). https://doi.org/10.4000/jtei.1433.

[20] Steuckardt, Agnès. *Entre village et tranchées. L'écriture de Poilus ordinaires*. Uzès: Inclinaison, 2015.

[21] Testa, Enrico. *L'italiano nascosto*. Torino: Einaudi, 2014.

[22] Thiers, Ghjacumu. «Aspects de la francisation au XIXème siècle, en Corse». *Études corse* 9 (1978): 5–39.

[23] Tomasi, Francesca. «XML/TEI per la trascrizione delle fonti primarie e la codifica dell'apparato critico». *Journal of Latin Linguistics* 9, fasc. 3 (2007): 129–48. https://doi.org/10.1515/joll.2007.9.3.129.

[24] Walter, Richard, Claire Bustarret, Marie Dupond, Alexandre Guilbaud, Giancarlo Luxardo, Yvan Leclerc, Jean-Sébastien Macke, Irène Passeron, Nicolas Rieucau, e Fabienne Vial-Bonacci. «L'Édition numérique de correspondances», 2018. https://cahier.hypotheses.org/files/2018/03/Correspondance_CAHIER.pdf.

[25] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Il progetto “MaximHum”: Italia umanistica e Moscovia cinquecentesca dialogano in digitale

Francesca Romoli[1], Letizia Ricci[2], Angelo Mario Del Grosso[3]
[1] Università di Pisa, Italia - francesca.romoli@unipi.it
[2] Università di Pisa, Italia - letizia.ricci@fileli.unipi.it
[3] CNR Istituto di Linguistica Computazionale, Italia - angelo.delgrosso@ilc.cnr.it

ABSTRACT[1]

Si presenta qui il progetto PRIN 2022 PNNR “Humanistic Italy and sixteenth-century Muscovy in dialogue: Digitization and digital mapping of the work of Maximus the Greek” – MaximHum (P2022837KN) che ha per scopo la digitalizzazione dell’opera edita di Massimo il Greco (1470ca-1556/1557) e la mappatura digitale della stessa. Attraverso l’individuazione di fenomeni notevoli, il lavoro di mappatura si propone in particolare di far emergere dai testi, a vari livelli (fattuale, letterario, concettuale), il ruolo di mediatore della cultura umanistica che Massimo il Greco – un greco bizantino formatosi in Italia e successivamente monacatosi sul monte Athos – svolse, lui per primo, nella Moscovia di Vasilij III e Ivan IV, pagando per l’azione riformistica che aveva promosso nel senso della modernità il prezzo di una lunga e severa reclusione. In questa sede si presenta il progetto e se ne discutono il contesto e la prima fase di digitalizzazione descrivendo l’ipotesi di lavoro per la creazione dell’archivio digitale dei testi e il piano del lavoro così come è stato messo a punto.

PAROLE CHIAVE

Massimo il Greco; Umanesimo; mediazione culturale; filologia digitale; archivi digitali.

1. INTRODUZIONE

Massimo il Greco, al secolo Michele Trivolis (Arta 1470ca - Sergiev Posad 1556/1557), visse in un tempo di cambiamenti epocali – la riunione delle Chiese al concilio di Ferrara-Firenze (1438-1439), la conquista ottomana dell’Impero bizantino (1453) e la *translatio imperii* da Roma a Mosca – e intersecò con il suo itinerario di vita i luoghi che di tali cambiamenti furono teatro: l’Impero sotto assedio dove nacque (1470ca), l’Italia umanistica dove si formò e dove lavorò come copista (1492-1506), il monte Athos (1506-1516) dove prese i voti monastici e il gran principato di Mosca dove si trasferì per porsi al servizio delle autorità in qualità di uomo di lettere e di monaco erudito e dove visse fino alla fine dei suoi giorni (1518-1556/1557). Come copista prima, negli ambienti della diaspora greca in Occidente, e come dotto poi, nei palazzi del potere di Mosca, Massimo il Greco partecipò attivamente ora alla trasmissione dell’eredità culturale greco-bizantina all’Occidente, ora al trasferimento in Moscovia dei valori dell’Umanesimo cristiano che in tale eredità trovavano il loro fondamento. In Moscovia, Massimo il Greco espresse ampiamente le sue abilità di umanista, facendosi primo tramite dell’Umanesimo e tentando di riformare l’ortodossia nello spirito della *renovatio christiana* occidentale. Se nell’immediato la sua attività gli valse due processi e una lunga e severa reclusione, in prospettiva l’opera di mediazione culturale di cui si rese protagonista creò un terreno di valori condivisi sul quale si sarebbero facilmente innestate sia la ‘Prima occidentalizzazione’ seicentesca, sia la ‘Grande occidentalizzazione’ promossa nel Settecento da Pietro il Grande. Gli studi sulla vita e sull’opera di Massimo il Greco vantano su scala internazionale una tradizione solida, ampia e diversificata quanto a obiettivi e metodi. In questo ambito, tuttavia, non è stato finora intrapreso alcun tentativo sistematico di elaborazione digitale. Il progetto PRIN 2022 PNNR “Humanistic Italy and sixteenth-century Muscovy in dialogue: Digitization and digital mapping of the work of Maximus the Greek” – MaximHum (P2022837KN) si propone di colmare questa lacuna con l’obiettivo particolare di portare a evidenza il ruolo di Massimo il Greco quale mediatore di e tra culture attraverso la digitalizzazione e la mappatura digitale della sua opera edita e delle sue traduzioni nelle lingue moderne. Il progetto si avvale della partecipazione di tre università italiane e della collaborazione dell’Istituto di Linguistica Computazionale “A. Zampolli” del CNR di Pisa (CNR-ILC). L’unità capofila è quella dell’Università di Pisa, guidata da Francesca Romoli, coordinatrice nazionale del progetto. Le unità locali sono dislocate alle Università di Bologna e di Chieti-Pescara e si trovano rispettivamente sotto la responsabilità di Alberto Alberti e di Maria Chiara Ferro. La

[1] La stesura del contributo è stato un lavoro collaborativo tra tutti gli autori. Francesca Romoli ha condotto la raccolta dati e ha redatto la Sezione 1 e la Sezione 2. Angelo Mario Del Grosso ha condotto l'analisi preliminare per la digitalizzazione e ha redatto la Sezione 3. Letizia Ricci ha impostato il primo tentativo di codifica e ha redatto la Sezione 4. L'abstract, le conclusioni e la rassegna bibliografica sono stati redatti congiuntamente da tutti gli autori. Tutti gli autori hanno contribuito significativamente alla revisione critica del contributo.

realizzazione del progetto prevede diverse fasi di lavorazione seguendo le migliori prassi di progetti simili di archivi digitali quali per esempio gli archivi latini (ALIM [7], MQDQ [9], DigilibLT [5]) e progetti tematici quali "Voci della Grande Guerra" [6]. La gamma delle attività va dal censimento dell'opera edita di Massimo il Greco e delle sue traduzioni nelle lingue moderne, alla creazione di una bibliografia aggiornata degli studi sul tema, alla trasformazione dei testi in formato digitale secondo quanto stabilito dalle linee guida TEI[2] [8], alla loro mappatura digitale, alla ricerca nell'ambito delle fonti (storico-documentarie, letterarie, linguistiche, bibliche) orientata dai dati emersi dal lavoro di mappatura digitale, fino all'elaborazione finale dei materiali e dei dati con la creazione di un sito dedicato. In ultimo, i dati saranno depositati in *repository* gestiti da infrastrutture di ricerca che ne assicurino l'aderenza alla progettazione e creazione di risorse FAIR [10] e che garantiscono la conservazione a lungo termine dei prodotti digitali della ricerca umanistica (si veda ad esempio l'infrastruttura H2IOSC). Il lavoro di censimento, preliminare alla trasformazione digitale dei testi, si pone come obiettivo la composizione di un inventario per quanto possibile completo della produzione di Massimo il Greco. Per ogni opera si registrano il titolo, la lingua, l'*incipit* e il *desinit*, l'edizione di riferimento del progetto e le edizioni esistenti, la natura del testo, segnalando cioè se si tratta di un componimento originale, di un'opera di traduzione, di un caso di 'riuso' o di un caso dubbio, eventuali dubbi di attribuzione, l'etichetta letteraria, la datazione, l'ambito tematico e la presenza di traduzioni. In inventario separato si raccolgono i manoscritti base e di controllo dell'edizione di riferimento del progetto per ogni opera, assegnando a ognuno un identificativo. In ulteriore inventario si catalogano le traduzioni delle opere, indicando la lingua di partenza e quella di arrivo, i necessari riferimenti bibliografici, ovvero l'edizione sulla cui base è stata eseguita la traduzione e gli estremi dell'edizione della traduzione, ed evidenziando la coincidenza o non coincidenza dell'edizione alla base della traduzione con l'edizione di riferimento del progetto. Il lavoro di mappatura digitale è finalizzato a rendere tangibile la componente umanistica dell'opera di Massimo il Greco e dunque manifesto il suo ruolo di testimone dell'Umanesimo e di primo tramite e mediatore della cultura umanistica nella Moscovia cinquecentesca. A questo scopo i testi saranno scandagliati, anche con l'ausilio di tecniche automatiche, alla ricerca di riferimenti a personaggi e luoghi dell'Italia umanistica, riferimenti agli esponenti dell'Umanesimo e citazioni dalle loro opere, riferimenti alle *auctoritates* e alle fonti della cultura umanistica e citazioni da esse, riferimenti ai concetti della cultura umanistica, del dibattito religioso e della polemica interconfessionale dell'epoca, occorrenze di citazioni e riferimenti biblici. I dati che emergeranno dalla mappatura saranno annotati nei testi mediante appositi marcatori TEI e raccolti in specifici indici così da renderli accessibili con ricerca guidata, per chiave semantica e con ricerca libera. Gli indici potranno riferirsi in parallelo a documenti di archivio, manoscritti, edizioni di testi per i quali sarà stato possibile ipotizzare o accertare un nesso con l'opera di Massimo il Greco e al thesaurus che sarà allestito sulla base dell'indice dei concetti. In questa sede si discuterà la prima fase di lavoro informatico del progetto, illustrando l'ipotesi di workflow che è stata messa a punto. Per meglio rappresentare il lavoro di codifica dei testi si porteranno esempi concreti delle fonti primarie da elaborare.

2. DESCRIZIONE DELLE FONTI

L'opera di Massimo il Greco è ampia, variegata e cronologicamente differenziata. La produzione in lingua slava, che ne rappresenta la componente maggioritaria, appartiene a due diversi periodi (dal 1518 al 1525 e dopo il 1531); a essa si aggiungono rare testimonianze in lingua greca relative agli anni del soggiorno in Italia e sull'Athos. I testi sono stati oggetto di due diversi progetti di edizione: il primo, che ha accompagnato ed espresso la riscoperta ottocentesca di Massimo il Greco, e l'altro, ispirato a più moderni criteri di scientificità, che è nato in seguito a un risvegliato interesse per la sua figura suscitato dalla sua canonizzazione (1988). Il secondo progetto di edizione è ancora in corso con la preparazione del terzo e ultimo volume. A questi progetti se ne sono aggiunti negli anni altri di maggiore o minore respiro, che hanno integrato la quantità dei testi editi disponibili. Nell'ambito del progetto si attingerà dunque a edizioni cronologicamente, tipologicamente e qualitativamente disomogenee, che spaziano dalle moderne edizioni critiche con apparato singolo (con annotazione delle varianti testuali) o con doppio apparato (con annotazione delle varianti testuali e con note descrittivo-esegetiche al testo e indicazione delle mende ivi apportate), alle edizioni più datate, che sono generalmente prive di apparato e si discostano dagli standard anche grafici delle edizioni più recenti. Neppure i formati delle fonti sono omogenei, essendo disponibili a seconda dell'epoca dell'edizione e delle politiche editoriali ora formati digitali testuali (per esempio PDF e DjVu), ora formati fotografici (perlopiù JPEG, TIFF), ora, infine, il solo formato cartaceo. Le principali edizioni di riferimento, che si trovano di seguito elencate in ordine cronologico e descritte quanto a consistenza e formato, saranno dunque diversamente rilevanti al processo di digitalizzazione:

- Bulanin, Dmitrij M. *Perevody i poslanija Maksima Greka. Neizdannye teksty*. Leningrad: Nauka, 1984: edizione critica di testi in lingua slava ecclesiastica con doppio apparato a piè di pagina (commenti al testo nelle note alfabetiche e

[2] TEI Consortium, eds. *Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/P5/

annotazione delle varianti testuali nelle note numeriche); pp. 276 (complessive di introduzione, indici, bibliografia, appendici); nr. testi: 19; formato: cartaceo [1].

- Grek, Maksim. *Sočinenija prepodobnogo Maksima Greka*, I-III. Kazan': Tipo-litografija Imperatorskogo Universiteta, 1894-1897 (1859-1862, 1 ed.): edizione di testi in lingua slava ecclesiastica priva di apparato, con note a piè di pagina in cui si indicano perlopiù i *loci* biblici citati nel testo; pp. 435+453+234; nr. testi: 28+49+53; formato: DjVu [3].
- Grek, Maksim. *Sočinenija prepodobnogo Maksima Greka v russkom perevode*, I-III. Svjato-Troickaja Sergieva Lavra: sobstvennaja tipografija, 1910-1911: traduzione russa dell'edizione precedente, con sporadiche note esplicative a piè di pagina; pp. 287+332+191; nr. testi: 49+28+53; formato: PDF non ricercabile [4].
- Grek, Maksim, prepodobnyj. *Sočinenija*, I-II. A a cura di Nina V. Sinicyna. Moskva: Indrik, Rukopisnye pamjatniki Drevnej Rusi, 2008, 2014: edizione critica di testi in lingua slava ecclesiastica con doppio apparato a piè di pagina (commenti al testo nelle note alfabetiche e annotazione delle varianti testuali nelle note numeriche); pp. 567+430 (complessive di introduzione, indici, bibliografia, appendici); nr. testi 23+51; formato: PDF ricercabile [2].
- Žurova, Ljudmila I. *Avtorskij tekst Maksima Greka: rukopisnaja i literaturnaja tradicii*, II. Novosibirsk: Izdatel'stvo Sibirskogo otdelenija Rossijskoj akademij nauk, 2011: edizione critica di testi in lingua slava ecclesiastica con doppio apparato a piè di pagina (commenti al testo nelle note alfabetiche e annotazione delle varianti testuali nelle note numeriche) e con esplicitazione dei loci biblici nelle note di chiusura di ogni testo (richiamate da asterischi); pp. 301 (complessive di introduzione, indici, bibliografia); nr. testi: 35; formato: cartaceo [11].

Di seguito si riportano, a titolo d'esempio, due immagini relative a due pagine prototipiche selezionate tra le tante possibili presenti nelle edizioni di riferimento. L'immagine in Figura 1a riporta una pagina dell'edizione a cura di Nina V. Sinicyna (2008). Si tratta di una edizione critica con duplice livello di apparato. Il primo, con esponenti alfabetici, contiene commenti al testo; il secondo, con esponenti numerici, le varianti testuali. In dettaglio, le regioni d'interesse richiamano: (a) La regione in alto con il riferimento alla pagina dell'edizione. Il dato è trascurabile per gli scopi del progetto. (b) Le "label", evidenziate in color arancio, л. 140об e л. 141 che indicano la paginazione del ms. di riferimento, ovvero il f. 141v e il f. 141r. Talora queste indicazioni si trovano integrate nel flusso del testo piuttosto che a margine. Per gli scopi del progetto dette informazioni possono essere rappresentate digitalmente e visualizzate poi a margine oppure integrate nel testo per mezzo di specifiche convenzioni editoriali. (c) Nei riquadri in blu sono collocati gli apparati. La sequenza di lettere indica i manoscritti nei quali si legge la variante riportata in nota. La legenda dei manoscritti è offerta prima di ogni testo. La rappresentazione delle informazioni di apparato è una delle attività più importanti e interessanti del lavoro di digitalizzazione. Di contro, l'immagine in Figura 1b, mostra una pagina dell'edizione stampata a Kazan' (1862). Dall'immagine si possono rilevare: (1) il numero in sequenza del testo; (2) il titolo del testo; (3) le note a piè di pagina, in cui si indicano perlopiù i *loci* biblici in occorrenza nel testo.

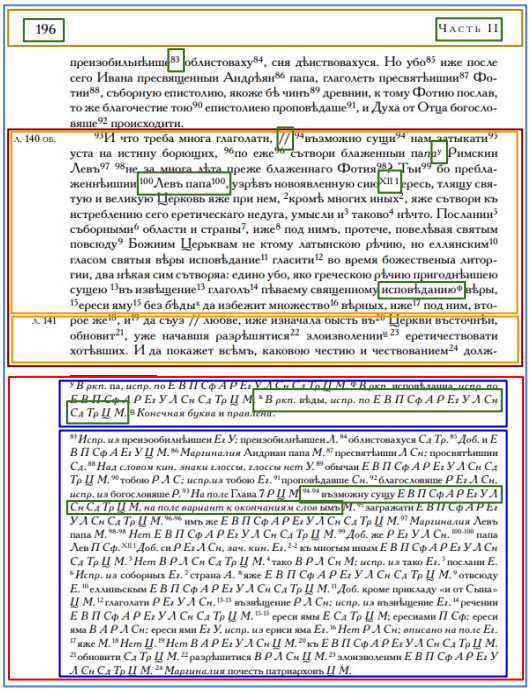


Figura 1a. Testo nell'edizione Maksim Grek. Sočinenija*, edizione a cura di Nina V. Sinicyna (2008)*

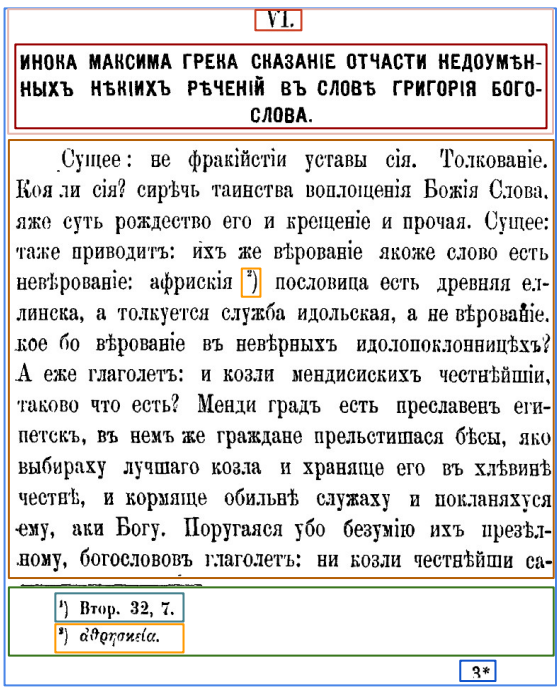


*Figura 1b. Testo nell'edizione Maksim Grek. S*očinenija prepodobnogo Maksima Greka*, III. Kazan' (1897)*

3. METODOLOGIA E WORKFLOW

Le prime fasi del progetto sono da un lato volte alla definizione del workflow dedicato alla creazione del repertorio dei testi digitali e, dall'altro, orientate alla definizione delle tipologie di fenomeni notevoli rilevanti. Rispetto a quest'ultimo, i dati d'interesse per l'indagine filologico-letteraria supportata da sistemi digitali prevede: 1) riferimenti a personaggi e luoghi dell'Italia umanistica; 2) riferimenti agli esponenti dell'Umanesimo; 3) citazioni alle opere di esponenti dell'Umanesimo; 4) riferimenti alle *auctoritates* e alle fonti della cultura umanistica ed all'esplicitazione delle citazioni da esse; 5) riferimenti ai concetti della cultura umanistica, del dibattito religioso e della polemica interconfessionale dell'epoca; 6) occorrenze di citazioni bibliche; 7) associazioni e nessi tra testi, elenchi e letteratura. Le diverse tipologie di dati emersi possono essere meglio prefigurate attraverso alcuni esempi.

Testo originale	Testo in traduzione
Флоренция* град есть прекраснѣишы и предобрѣиши сущих въ Италии градовъ, ихже азъ видѣх. В том градѣ манастырь есть, мниховъ отчина, глаголемых по-латыньскы предикаторовъ […] Храм же священыя сея обители святѣишаго апостола и еуаггелиста Марка** получивъ призирателя и предстателя.	Firenze è di gran lunga la città più bella e più salubre di tutte le città italiane che ho visto. In quella città c'è un convento che è dimora dei frati che in latino sono chiamati *predikatori* […] La chiesa di questa santa casa ha eletto a suo custode e protettore il santissimo apostolo ed evangelista Marco.

Tabella 1. Testo Parallelo di Maksim Grek in forma full-text da analizzare e annotare

Il testo riportato in Tabella 1 con elencati tre diversi livelli di rappresentazione dei dati di rilievo: 1) livello della rappresentazione del testo, 2) livello dell'annotazione/arricchimento semantico e funzionale, e 3) livello di espansione e approfondimento delle informazioni.

- Livello 1: rappresentazione del testo. Risorsa rappresentabile mediante l'adozione di elementi strutturali definiti in vocabolari standard, quali ad esempio XML/TEI, funzionali alla costituzione del corpus parallelo (testo originale e dove presente traduzione). Ad esempio, gli anonymous block (`<ab>`), porzioni a granularità più fine (`<seg>`), nonché selezioni di testo semanticamente e funzionalmente connotate, facendo uso anche di milestone, ancore e annotazione in stand-off.
- Livello 2: annotazione. Luoghi, persone, organizzazioni, oggetti presenti in città, rappresentabili mediante opportuni elementi XML/TEI per le entità nominate quali `<name>`, `<placeName>`, `<orgName>`, `<objectName>` da rimandare a descrizioni canoniche oppure a liste di autorità. Nell'esempio di cui sopra, "Firenze" è un nome di luogo, mentre "Convento" si riferisce al "Convento domenicano di San Marco" e potrebbe essere annotato come organizzazione oppure come oggetto. In entrambi i casi una dettagliata descrizione potrebbe prevedere l'uso dell'elemento `<location>` per registrarne le coordinate geografiche così come il nome del luogo geo-politico dove è situato
- Livello 3: espansione. Associazioni e espansioni mediante: 1) Immagini. Nel caso dell'esempio immagine di Firenze e/o del convento di San Marco, di epoca storica o contemporanee. 2) Nessi con gli indici delle entità nominate. Per esempio l'associazione ai personaggi della Firenze quattrocentesca citati nell'opera di Massimo il Greco, quali Girolamo Savonarola. 3) Nessi con gli elenchi delle *auctoritates* e delle fonti della cultura umanistica quali per esempio collegamenti ai manoscritti copiati da Massimo il Greco a Firenze (i.e., Biblioteca Medicea Laurenziana, Conv.soppr. 104, Pseudo Dionigi Areopagita, *De divinis nominibus* - possibile link a BML)

Le associazioni, espansioni, nessi, ed altri approfondimenti sono registrati mediante il modello Web Annotation Data Model all'interno del blocco stand-off del vocabolario TEI a creare un grafo di navigazione tra le entità messe così in relazione. Sulla base dei requisiti esposti, il workflow di rappresentazione digitale e di elaborazione computazionale dei testi prevede una pianificazione in diverse fasi. Il processo si differenzia in base ai differenti formati delle fonti. La prima fase riguarda l'attività di acquisizione dei materiali cartacei per mezzo di riproduzioni fotografiche facsimilari. La valutazione iniziale si è soffermata dunque sulle modalità da utilizzare per la raccolta dei materiali e per la loro descrizione mediante opportuna *metadatazione* e inventariazione. Funzionale alla fase di acquisizione è l'utilizzo di macchine fotografiche ovvero di scanner planetari, normalmente utilizzati per la digitalizzazione di libri antichi e manoscritti. Le soluzioni individuate sperimentano un costo elevato, necessitano di personale con preparazione specifica e di tempi di attesa spesso troppo dilatati. Cercando il miglior compromeso tra qualità e costo, il processo di acquisizione è stato orientato sulle capacità dello scanner disponibile nel Laboratorio di Cultura Digitale dell'Università di Pisa. Si tratta di uno scanner standard planetario senza contatto con la fonte, utile dunque per salvaguardare l'integrità del supporto. Inoltre, lo strumento

in dotazione prevede un processo di acquisizione dell'immagine semplice e veloce. La risoluzione delle immagini ottenute così come eventuali strumenti di post-processing, quali *ScanTailor*, sono indispensabili per l'efficacia delle fasi successive di riconoscimento del testo. Il processo infatti segue con la fase relativa al riconoscimento automatico del testo in formato digitale full-text. Ad oggi contiamo su diverse applicazioni sia commerciali sia open source per il riconoscimento ottico dei caratteri (OCR). Ad esempio strumenti quali *Abbyy FineReader*, *Tesseract*, *Kraken*, *NAPS2*, sono stati valutati rispetto all'accuratezza del risultato ottenuto a partire dalle immagini disponibili. La tecnologia OCR applicata ad edizioni critiche di interesse umanistico in lingua slava ha alcune peculiarità. Nello specifico, il lavoro di acquisizione si basa soprattutto su edizioni critiche con singolo e, in alcuni casi, doppio registro di apparato. I sistemi di OCR presentano in prima battuta alcune limitazioni al riconoscimento di specchi testuali con divisione strutturata e molteplici flussi testuali, come il rapporto topografico tra testo e apparato critico. Spesse volte, infatti, tali applicazioni uniscono i diversi flussi testuali in un unico flusso di testo, alterando i riferimenti alle note. Inoltre, le varietà linguistiche dei testi possono essere assenti nel dizionario di riferimento se non opportunamente personalizzato secondo variabili diacroniche.
Tale mancanza impatta sul risultato finale poiché il riconoscimento potrebbe tralasciare set di caratteri non presenti nell'addestramento sull'alfabeto cirillico; in particolare grafemi espunti dal cirillico russo con la riforma ortografica del 1918 presenti in alcune edizioni del progetto.
In Tabella 2 un esempio tratto dai primi test condotti su tre sistemi OCR applicati allo stesso campione di testo. Una volta ottenuti i dati in formato testuale, la terza fase prevede quindi la modellazione dello schema di codifica per la rappresentazione strutturata delle risorse. A tale fase corrisponde un'opportuna personalizzazione del vocabolario XML/TEI mediante la definizione di un documento ODD. Lo schema di codifica, come atteso, seguirà le ipotesi e i requisiti di rappresentazione descritti nei livelli di codifica precedentemente introdotti.

Abbyy FineReader	Tesseract	NAPS2
Времл, кђ которому относитсл просв-ћтителБнан дбнтелБностБ вђ Росс1и преп. Максима Грека, бнло временемЪ разнмхЪ печалвннхЪ нестроенjб, происхо-дившихђ вђ неи частjго ОТЪ разннхЪ иноземннхЪ вл1'л- Н1и, частJК) отђ домашнихЪ внутреннихЪ причинЪ.	Время, къ которому относитол просвѣтительная дЪятельность въ Россїи преп. Максима Грека, было временемъ разныхъ печальныхъ нестроеній, происхо- дившихъ въ ней частію отъ разныхъ иноземныхъ влія- ній, частію отъ домашнихъ внутреннихъ причинъ.	Время, къ которому относится просвтительная дЪятельность въ Росси преп. Максима Грека, было временемъ разныхъ печальныхъ нестроенй, происхо- дившихъЪ въ ней часто отъ разныхъ иноземныхЪ вл1я- н!й, частю отъ домашнихъ внутреннихъ причинъ.

Tabella 2. Estratto dei risultati del processo di Riconoscimento Ottico dei Caratteri applicato ai testi di riferimento

4. MODELLO DEL TESTO E SCHEMA DI CODIFICA

La ricchezza e la varietà dei testi ci suggeriscono la definizione di uno schema di codifica che consideri le caratteristiche comuni a tutti i materiali e raffinarlo sulla base delle nuove esigenze adottando una metodologia iterativa durante il lavoro di codifica. I documenti digitali delle edizioni sono codificati in singole unità testuali e, successivamente, saranno collegati dinamicamente in un unico TEI corpus, necessario per la pubblicazione e per la fruizione dell'archivio digitale. Per identificare univocamente i testi codificati, l'elemento radice `<TEI>` di ciascun file ha un attributo `@xml:id` il cui valore deriva dalla classificazione dei testi avanzata in Ivanov (1969).
Il `<teiHeader>` comprende i metadati di ciascuna opera seguendo l'inventario già predisposto. L'elemento `<fileDesc>` contiene una descrizione bibliografica del file elettronico, tra gli elementi annidati, `<sourceDesc>`, indica, sottoforma di riferimento bibliografico, le edizioni in cui è presente il testo e la posizione al loro interno. Si distingue l'edizione di riferimento da cui è tratto il testo da altre edizioni in cui è presente lo stesso. In `<sourceDesc>`, inoltre, sono registrate le informazioni relative ai manoscritti: il manoscritto base dell'edizione di riferimento negli elementi descrittivi `<msIdentifier>` e `<msContent>` di `<msDesc>` e la lista dei testimoni di controllo, `<listWit>`, collegata all'apparato. L'elemento `<profileDesc>` fornisce la descrizione degli aspetti non prettamente bibliografici, quali la datazione, la lingua e la natura del testo.
La struttura dell'elemento `<text>` prevede il tag `<body>` al cui interno è presente il testo pericopato suddiviso in elementi `<seg>`, mentre in `<back>` si registrato i riferimenti di apparato. Le informazioni peritestuali relative al manoscritto, quali interruzione di pagina e numerazione del foglio, sono espresse sul testo rispettivamente con `<metamark>`, che descrive

il segno grafico (//) con funzione “end-page”, e <pb> per l’indicazione del foglio presente a margine o internamente al testo. Per quanto riguarda la marcatura delle notizie di apparato è stato scelto, tra i tre metodi proposti dalla TEI, il *double end-point attachment* che consente un riferimento preciso tra la porzione di testo e l’indicazione di apparato, consentendo anche la gestione di porzioni testuali sovrapposte: in questo modo è possibile ottenere una riproduzione fedele della struttura dell’edizione di riferimento cartacea. Tuttavia ciascuna edizione critica usa sistemi di riferimento differenti per quanto riguarda la numerazione delle entrate d’apparato, la distinzione in singolo o doppio apparato e la collocazione dei riferimenti a citazioni bibliche. L’edizione critica digitale ha lo scopo di creare un’unica struttura coerente. Gli apici numerici e letterari inseriti nel testo, che rimandano rispettivamente all’apparato e alle note descrittive, sono espressi dal tag <anchor> che attribuisce un punto di ancoraggio alla parola o alla porzione di testo attraverso un @xml:id collegato alla specifica entrata di apparato o nota nel <back>. L’apparato critico è organizzato sfruttando l’elemento <listApp> (vd. Fig. 2) in cui sono indicate le varianti riportate dai diversi testimoni. L’elemento <app> indica l’entrata d’apparato con la lezione ed una o più varianti, <rdg>, con l’indicazione del testimone che veicola la lettura nell’attributo @wit, che a sua volta rimanda alla lista dei testimoni.

```
<listApp>
   <head rend="super">1-1</head>
   <app from="#v1a1" to="#v1a2">
      <witDetail wit="#Сл">Нет</witDetail>
   </app>
</listApp>
<listApp>
   <head rend="super">2</head>
   <app from="#v2">
      <rdg wit="#П">како</rdg>
   </app>
</listApp>
```

Figura 2. Snippet di codifica dell’apparato nel testo Ivanov 332 (Žurova 2011)

Le note di tipo descrittivo-esegetico sono espresse in <noteGrp>, distinte dall’apparato critico perché trattasi di commenti interpretativi sul testo. In casi particolari in cui la nota è propriamente una lezione - in quanto precisa la lettura di un testimone accolta dall’editore sul testo e riposa in nota come variante la lettura del manoscritto base dell’edizione - si è deciso di isolarla e non trattarla come semplice nota, ma strutturare l’informazione in un <listApp>, distinto dall’apparato critico, in cui la lettura <rdg> nell’attributo @wit rimanda al manoscritto base. I riferimenti delle citazioni bibliche nelle edizioni non sono sempre espressi allo stesso modo, infatti possono essere inclusi nelle note descrittive oppure essere riportate in coda al testo con l’indicazione del foglio. Nella codifica di queste informazioni si è deciso di raggruppare le citazioni in <listBibl> e per ogni <bibl> inserire un <ref> che rimanda alla citazione sul testo strutturata con l’elemento <cit> e <quote xml:id="...">. Questo ci consente di isolare tutti i rimandi biblici, ma anche di distinguere citazioni bibliche da altri tipi di citazioni quando presenti.

5. CONCLUSIONI

Il contributo ha introdotto le prime fasi del progetto PRIN 2022 PNNR “Humanistic Italy and sixteenth-century Muscovy in dialogue: Digitization and digital mapping of the work of Maximus the Greek” – MaximHum (P2022837KN). L’idea che si propone per la prossima presentazione del corpus in ambiente Web è quella di avere a disposizione la lista dei testi e per ciascuno la possibilità di visualizzare contestualmente i metadati di riferimento e su scelta dell’utente le varianti di apparato, le note descrittive e le citazioni bibliografiche cliccando sui rispettivi termini evidenziati nel testo. In seconda battuta si andrà a sviluppare sul sito una sezione di ricerca in cui l’utente può individuare i testi d’interesse secondo filtri di ricerca. La ricerca potrà essere condotta ad esempio in base a criteri di metadatazione quali testi ricercati per edizione di riferimento, manoscritto base, etichetta letteraria, oppure per informazioni testuali come riferimenti biblici, parola chiave che rimanda a termini notevoli sul testo (persone, luoghi, concetti).

6. RINGRAZIAMENTI

Il progetto PRIN 2022 PNRR “MaximHum” è stato realizzato grazie ai finanziamenti D.D. nr. 1234 dell’1.08.2023.

BIBLIOGRAFIA

[1] Bulanin, Dmitrij M. *Perevody i poslanija Maksima Greka. Neizdannye teksty*. Leningrad: Nauka, 1984.

[2] Grek, Maksim. *Sočinenija, I-II*. A cura di Nina V. Sinicyna. Moskva: Indrik, Rukopisnye pamjatniki Drevnej Rusi, 2014.

[3] Grek, Maksim. *Sočinenija prepodobnogo Maksima Greka, I-III. Kazan'*. I. Tipo-litografija Imperatorskogo Universiteta, 1894.

[4] Grek, Maksim. *Sočinenija prepodobnogo Maksima Greka v russkom perevode, I-III*. Svjato-Troickaja Sergieva Lavra: sobstvennaja tipografija, 1910.

[5] Lana, Maurizio. «Metodologie e problematiche per una biblioteca digitale. Il Caso Di digilibLT». *DigItalia* 7, fasc. 1 (2012): 40–64.

[6] Lenci, Alessandro, Nicola Labanca, Claudio Marazzini, Simonetta Montemagni, Federico Boschetti, Irene De Felice, Stefano Dei Rossi, et al. «Voci della Grande Guerra. Preserving the Digital Memory of World War I». In *Patrimoni Culturali Nell'era Digitale. Memorie, Culture Umanistiche e Tecnologia. AIUCD2018 - Book of Abstracts*, a cura di Daria Spampinato, 193–95. Bari: Quaderni Di Umanistica Digitale, 2018. https://doi.org/10.6092/unibo/amsacta/5997.

[7] Russo, Luigi. «ALIM, Archivio della latinità italiana del Medioevo ALIM, Archivio della latinità italiana del Medioevo». *Reti Medievali Rivista* 6, fasc. 1 (2005): 149–51. https://doi.org/10.6092/1593-2214/181.

[8] Tei Consortium, (a cura di). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.7.0*. Tei Consortium, 2024. http://www.tei-c.org/Guidelines/P5/.

[9] Venuti, Martina, Angelo Mario Del Grosso, Federico Boschetti, Luigi Tessarolo, Alessia Prontera, Dylan Bovet, Gianmario Cattaneo, e Melis Melis. «La 'Galassia MQDQ': un concetto di filologia tradizionale, digitale, sostenibile». *Relations* 4, fasc. 1 (2023): 71–120. https://doi.org/10.30687/mag/2724-3923/2023/07/003.

[10] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[11] Žurova, Ljudmila I. *Avtorskij tekst Maksima Greka: rukopisnaja i literaturnaja tradicii, I-II*. Novosibirsk: Izdatel'stvo Sibirskogo otdelenija Rossijskoj akademij nauk, 2008.

L'Archivio di Giuseppe Fava: conservazione e valorizzazione attraverso il digitale

Giuseppe Davide Di Mauro[1], Marzia D'Amico[2]
[1] Università di Catania, Italia - giuseppe.d.dimauro@gmail.com
[2] Università di Catania, Italia - damico.marzia96@gmail.com

ABSTRACT

Ricordato principalmente come giornalista assassinato dalla mafia, Giuseppe Fava è stato anche un artista e un intellettuale poliedrico. A seguito della sua morte, la figlia Elena si è dedicata alla raccolta di tutto il materiale relativo all'opera del padre, dando successivamente vita a questo scopo a una Fondazione che ne porta il nome. L'importante patrimonio culturale costituito attraverso questa iniziativa si presta a essere conservato e reso fruibile in modo durevole dagli strumenti offerti dal digitale. Il progetto qui illustrato si propone, di conseguenza, di perseguire entrambi questi scopi attraverso l'attenta modellizzazione della digitalizzazione dell'Archivio, ponendo particolare attenzione all'edizione dei testi che richiedono cura filologica, per mezzo della loro codifica in XML/TEI. L'obiettivo è quello di favorire il maggior grado di fruibilità possibile di questo complesso documentario e agevolare in tal modo la riscoperta e la piena comprensione della figura intellettuale dell'autore siciliano.

PAROLE CHIAVE

Giuseppe Fava; digital archives; XML/TEI encoding; digital philology.

1. INTRODUZIONE

Riscoprire Giuseppe Fava

Quando la mafia, il 5 gennaio 1984, ha assassinato Giuseppe Fava, ne ha reso eterna l'immagine di giornalista impegnato a svelare i legami tra criminalità organizzata, imprenditoria, politica e media. Parallelamente, però, questa forzata ascesa nel pantheon dei martiri dell'antimafia ha posto in ombra la vasta ed eterogenea produzione che era valsa all'autore, in vita, notorietà nazionale e, a tratti, europea.

Fava fu infatti anche prolifico drammaturgo, romanziere, pittore, autore e conduttore radiofonico. I suoi testi teatrali e narrativi hanno attirato gli interessi dei cineasti (due le trasposizioni dai suoi lavori), tanto che lo stesso Fava è stato chiamato a realizzare la sceneggiatura (poi sviluppata da lui stesso nel romanzo *Passione di Michele*) di *Palermo oder Wolfsburg* del regista Werner Schroeter, premiato con l'Orso d'oro per il miglior film al Festival di Berlino del 1980.

Grazie a tutto questo, proprio all'altezza del 1980, Fava sembrava avviato ad affermarsi nel panorama intellettuale nazionale. Quell'anno, però, segna un punto di svolta fondamentale per la sua vita; egli ricevette infatti la proposta di assumere la direzione del nascente 'Giornale del Sud', offerta per la quale lasciò Roma, dove si era trasferito da qualche anno, per tornare a Catania. L'impegno come direttore del nuovo quotidiano portò Fava a mettere quasi del tutto da parte l'attività artistica e, allo stesso tempo, ad avviare un'opera di smascheramento dei meccanismi del potere operanti nel capoluogo etneo che lo condusse alla rottura con i suoi editori e alla conseguente fondazione del mensile 'I Siciliani'. Le inchieste realizzate da Fava e dai suoi 'carusi' per questa rivista costituiscono il movente per il quale venne deciso e compiuto il suo assassinio.

È tale stretta connessione tra la sua produzione giornalistica e la sua tragica morte ad aver favorito da una parte la sua identificazione come giornalista ucciso dalla mafia e dall'altra, già dal fatidico 1984, una quasi immediata perdita di attenzione nei confronti della sua attività artistica. Essa, però, non possiede soltanto un'oggettiva qualità, ma risulta di grande interesse e rilevanza per lo stretto legame che intreccia con la parabola biografica e professionale del Fava giornalista, rivelandosi fondamentale non solo per un'esaustiva ricostruzione della sua figura intellettuale, ma anche per poter pienamente comprendere il pensiero di un uomo che, sulla rampa di lancio della fama e del successo, ha scelto consapevolmente la strada accidentata del giornalismo d'inchiesta in Sicilia, seguendo una forte pulsione etica e civile.

2. STATO DELL'ARTE

Soffocata dalla pur meritata attenzione per l'attività giornalistica, l'opera artistica di Fava è passata in secondo piano e di conseguenza pochi sono gli studi ad essa dedicati [1, 2, 3, 4, 9, 10, 11].

Sono invece numerosissimi gli interventi circa le conseguenze che l'utilizzo del nuovo medium digitale ha sulle pratiche filologiche e sono sempre di più i progetti di allestimento di archivi ed edizioni digitali [7]. Gli studiosi si sono trovati di

fronte a nuovi quesiti, si sono interrogati su cosa sia un «edizione scientifica digitale» [8], e una delle risposte più pregnanti è giunta da Patrick Sahle che sottolinea come un'edizione scientifica digitale sia tale se non può essere stampata senza perdita di informazioni [12]. Si è riflettuto su come debbano cambiare le competenze del filologo, e come cambi anche il rapporto che queste edizioni intrattengono con i loro lettori e fruitori [6, 8]. Nel tentativo di comprendere come vadano (ri-)definiti i termini 'archivio' ed 'edizione' nel contesto digitale, Francesca Tomasi ha proposto il termine «Knowledgesites», «Ambienti di conoscenza», individuando con questa definizione quelle risorse che, attraverso un approccio multidisciplinare, offrono agli utenti «un'esperienza conoscitiva completa» [14].
Ottimo esempio di *knowledge site* che coniuga un approccio filologico attento e strumenti utili a studiosi e studenti è il progetto Pirandello Nazionale[1], sviluppato presso l'Università di Catania. Il sito permette la consultazione dei testi dell'autore (di cui fornisce trascrizione delle testimonianze manoscritte, edizione critica statica e dinamica) e ad essi coniuga diverse risorse (tra cui percorsi didattici, risorse multimediali, vocabolari) [5, 13]. Di notevole interesse sono anche il portale Manzoni Online[2] e l'Edizione Nazionale delle Opere di Aldo Moro[3].

3. IL PROGETTO

L'Archivio

Della vasta produzione di Fava, e di molta rimasta allo stato di progetto, è raccolta un'ampia messe di documenti nell'Archivio della Fondazione Giuseppe Fava, costituita nel 2002 proprio allo scopo di conservare i materiali relativi all'opera dell'autore e di trasmetterne la memoria.
Fava era piuttosto metodico nell'organizzare la propria poliedrica attività, tanto che solitamente sistemava in prima persona, all'interno di carpette, tutto ciò che riguardava le varie opere. Alla sua morte, la figlia Elena ha riunito e trasferito nella propria abitazione i documenti del padre, comprendenti anche quelli relativi agli anni giovanili e della formazione e quelli audiovisivi legati alle collaborazioni di Fava con la radio, il cinema e la televisione. A partire dalla costituzione della Fondazione, e in misura sempre più strutturata a partire dal 2014, Elena Fava e il marito Giuseppe M. Andreozzi hanno provveduto a organizzare quanto raccolto per tipologia, strutturando di conseguenza l'Archivio nelle tre macroserie 'Formazione (1931-1947)', 'Prime stesure (1942/1943-1955)' e 'Attività (1945-1983)'. Quest'ultima è a sua volta articolata in sette sezioni: 'Teatro (1945-1983)', 'Racconti Narrativa Saggistica (1955-1981)', 'Giornalismo Produzione (1947-1983)', 'Radio (1977-1978)', 'Film (1979-1983)', 'Progetti di Attività (1968-1983)', 'Pittura e Grafica (1959-1983)'. Per intendere la mole di materiale conservato, basti considerare che le sole sezioni 'Teatro' e 'Racconti Narrativa Saggistica' comprendono rispettivamente 9196 e 5546 carte.
L'Archivio così articolato è stato dichiarato di interesse culturale con il decreto n.71 del 27 giugno 2018 dalla Soprintendenza Archivistica della Sicilia – Archivio di Stato di Palermo (MiBACT), al quale è seguita, nel 2021, la stipula di una convenzione tra la Fondazione e la Direzione Generale Archivi che prevede un nuovo ordinamento e l'inventariazione completa dei documenti, in corso di compilazione a opera dell'archivista Simone Lisi. Completata la riorganizzazione, il materiale sarà suddiviso in dieci serie: 'Formazione (1927-1947)', 'Scritti giovanili (1942/43-1951)', 'Teatro (1945-1983)', 'Giornalismo (1948-1983)', 'Narrativa, Saggistica (1955-1981)', 'Disegni (e Pittura e Grafica) (1959-1983)', 'Film (1971-1981)', 'Radio (1977-1978)', 'Progetti (1953-1983)', 'Raccolta Giornali d'epoca'.
La descrizione delle serie e delle unità archivistiche già inventariate (afferenti a 'Formazione', 'Prime stesure' e 'Teatro') è consultabile tramite l'applicazione web ArchiVista 3.1.1 dalla pagina relativa all'Archivio[4] sul sito internet della Fondazione Fava. Ai fini dello studio dell'Archivio, importante risulta che l'operazione di riordino prevede l'assegnazione di una nuova segnatura a tutti i documenti, anche a seguito della ricollocazione di alcuni di essi.

Proiettare l'Archivio nel digitale

Assodata la rilevanza culturale dell'Archivio di Giuseppe Fava e considerata la deperibilità dei materiali originali contenenti l'opera dell'autore, abbiamo maturato il proposito di elaborare un progetto di digitalizzazione di questa preziosa risorsa, allo scopo di poter a un tempo conservare i documenti nel loro stato di fatto e permettere la loro consultazione a una platea di possibili fruitori ben più ampia di quella che l'attuale collocazione fisica dell'Archivio stesso può consentire. Se già il proposito della conservazione rende auspicabile la digitalizzazione, è però sul secondo scopo che abbiamo considerato fondamentale incentrare la nostra proposta. Riteniamo, infatti, che solo consentendo il maggior numero di utilizzi possibile dell'Archivio si possa favorire la riscoperta di Giuseppe Fava. Per realizzare lo scopo, risulta necessario

[1] Consultabile al sito https://www.pirandellonazionale.it/
[2] https://www.alessandromanzoni.org/
[3] https://aldomorodigitale.unibo.it
[4] https://www.fondazionefava.it/archivio-di-giuseppe-fava/. Si ringrazia la Fondazione Giuseppe Fava e in particolare il Professore Giuseppe Maria Andreozzi per aver concesso la consultazione dei materiali dell'Archivio.

modellizzare [7] il progetto in modo preciso e funzionale e ciò comporta ipotizzare in che modo i documenti raccolti possano essere utilizzati e da quale tipo di fruitore. In relazione a questo aspetto, ma anche in risposta a criteri di economicità sia in fase di realizzazione che di gestione, è sorta la necessità di classificare i documenti in base alla metodologia più opportuna per digitalizzarli.
Limitandoci in questa sede al materiale legato all'ambito narrativo, teatrale e giornalistico (che costituisce il focus privilegiato del nostro progetto), le principali forme di utilizzo possono essere raggruppate nella sfera divulgativo-didattica, in quella scientifica e in quella documentaria. I materiali afferenti a quest'ultima funzione e che possono essere ritenuti 'accessori' (ad esempio le pagelle scolastiche o le brochure teatrali) non richiedono i massimi standard qualitativi per la loro riproduzione fotografica. Per tutto ciò che, invece, si presta ad utilizzi differenti, come ad esempio i testi narrativi e quelli teatrali, bisogna valutare in quali casi poter procedere con una digitalizzazione tramite programma di OCR, corretto manualmente (con l'eventuale indicazione di normalizzazioni e altre correzioni) e in quali invece si dovrà realizzare una vera e propria forma di edizione; in questo secondo caso, si renderebbe necessario riprodurre i documenti mediante fotografie in alta risoluzione e archiviare queste in formato TIFF.
Ci si rende dunque conto che, se buona parte del materiale di Fava si presta a una digitalizzazione relativamente rapida, una parte altrettanto corposa, se non maggioritaria, richiede un'attenzione di tipo filologico. Di conseguenza, altro scopo del nostro progetto è quello di produrre, per quei testi che lo richiedano, un'edizione interpretativa secondo standard XML/TEI che possa fornire da una parte allo studioso, anche tramite visualizzazione affiancata al facsimile mediante EVT [8], una base da cui partire per la realizzazione di studi filologici, dall'altra (dove possibile) al semplice lettore un testo fruibile. Per i casi più importanti, come i romanzi, o più complessi filologicamente è nelle intenzioni la realizzazione di vere e proprie edizioni critiche digitali [8, 12].
Il complesso di materiale digitale prodotto (rispondente a standard di condivisione IIIF per le riproduzioni fotografiche) potrebbe essere fruito in *open access* (almeno in parte, previa disponibilità della Fondazione e della famiglia Fava) in una sezione apposita all'interno del sito della Fondazione oppure strutturato in un portale indipendente, ma comunque dialogante con la sezione 'Archivio' del sito, ad esempio con l'inserimento dei link relativi alle singole risorse nella descrizione della scheda d'archivio corrispondente.

Un caso studio: *La ragazza che fu uccisa in luglio*

Per provare a mostrare quanto i documenti originali di Fava si prestino a studi di filologia digitale e, per converso, come gli strumenti dell'umanistica digitale possano rivelarsi particolarmente utili per lo studio delle opere dell'autore palazzolese, abbiamo scelto come esempio il caso della prima stesura conservata del racconto *La ragazza che fu uccisa in luglio*, testo significativo all'interno della produzione narrativa di Fava, per il riferimento a fatti storici reali (legati alla sua biografia), per il tema amore/morte tipico dell'autore e per l'interessante caso filologico che costituisce.
Qualche cenno su quest'ultimo aspetto: nel 1993, la piccola casa editrice Il Girasole pubblica *La ragazza di luglio*, un racconto di Fava conservato in Archivio in un dattiloscritto in pulito indicato con segnatura RNS10[5] e datato con qualche incertezza tra il 1958 e il 1960. Si tratta del racconto del nascente amore tra due giovani che finisce in tragedia a causa del bombardamento di Palazzolo Acreide avvenuto il 9 e 10 luglio 1943. Questo stesso nucleo narrativo, ma sviluppato in modo completamente diverso, ha un suo antecedente non collazionabile ne *La ragazza che fu uccisa in luglio*, racconto apparso nell'edizione del 18-19 febbraio 1957 di 'Espresso sera'[6]. A sua volta, di questo testo esiste una precedente redazione, datata tra 1951 e 1955, inserita in apertura di una raccolta di quattro racconti in totale, dattiloscritta e rilegata dallo stesso Fava in un fascicoletto recante in copertina un'illustrazione e il titolo *Le vergini del Sud*, probabilmente realizzato per essere proposto alle case editrici. Questa redazione, in diciassette carte e un frontespizio battuti a macchina sul solo fronte, è realizzata su carta molto sottile e trasparente ed è ricca di correzioni autografe, sottolineature e appunti anche sul verso dei fogli, redatti con penne di diversi colori. Il testo dattiloscritto è in buona parte sovrapponibile a quello apparso su 'Espresso sera', ma la quasi totalità delle correzioni apportate su di esso non si trovano né in quest'ultima versione né nel testo de *La ragazza di luglio*. Esse sembrerebbero dunque posteriori al febbraio del 1957 e la presenza di alcuni fogli di riuso dattiloscritti contenuti sciolti all'interno de *Le vergini del Sud*[7] e databili con certezza al 1963 farebbero pensare che intorno a quest'anno Fava abbia ripreso in mano i racconti della raccolta per rielaborarli. È possibile ipotizzare che, abbandonata l'idea della revisione, si sia dedicato alla scrittura di una nuova versione del racconto sul bombardamento

[5] Le segnature indicate sono quelle apposte ai documenti da Elena Fava e Giuseppe M. Andreozzi, poiché al materiale in esame non è stata ancora associata la nuova segnatura.
[6] Una bozza di stampa della pagina di giornale è conservata in Archivio con segnatura RNS09.
[7] Unità archivistica PS106. Di questo documento, nel nuovo ordinamento dell'Archivio, è previsto lo spostamento da 'Prime stesure' a 'Narrativa, Saggistica'.

di Palazzolo di cui l'edizione postuma de Il Girasole rappresenta l'approdo, ma della quale non si conserva alcuna stesura antecedente RNS10 (la cui datazione slitterebbe così in avanti).

4. LA CODIFICA XML/TEI

La codifica filologica

Per l'allestimento di un'edizione scientifica digitale interpretativa del testimone interno a PS106, che presenta una complessa situazione filologica, ci siamo serviti del modello di codifica TEI.

Tutte le informazioni relative al fascicolo e alle carte del racconto sono fornite nel `<teiHeader>`, in particolare all'interno del tag `<msDesc>` contenente: il tag `<msIdentifier>`, che fornisce le informazioni relative alla collocazione fisica delle carte; il tag `<msContents>`, che illustra brevemente il contenuto del fascicolo a cui le carte appartengono e infine il tag `<physDesc>`.

La stesura originaria, dattiloscritta, presenta svariati interventi autografi a penna. Le penne che l'autore utilizza sono numerose; di esse sono immediatamente distinguibili gli inchiostri di colore differente, ma è risultato impossibile distinguere tra inchiostri dello stesso colore. Abbiamo pertanto ritenuto opportuno limitarci, nella codifica, all'indicazione dei diversi colori, individuando in questo modo cinque penne: una di colore blu molto scuro (utilizzata per gli interventi immediati, come la correzione di refusi), una blu, una nera, una rossa e una verde. Di conseguenza, nel `<physDesc>` sono stati inseriti i tag: `<objectDesc>`, contenente la descrizione fisica del fascicolo; `<typeDesc>`, in cui un `@xml:id` denominato "Fava" indica la stesura base dattiloscritta; `<handDesc>`, all'interno del quale vengono definite le varie penne in base al colore dell'inchiostro (indicate da valori di `@xml:id` quali, "FavaB", "FavaN", ovvero una sigla composta dal nome dell'autore e dall'iniziale del colore della penna, oppure, nel caso di "FavaI", dalla natura degli interventi realizzati, appunto immediati) (vd. Fig. 1).

```
37            <physDesc>
38               <objectDesc>
39                  <p>Il racconto è dattiloscritto sul solo fronte di 18 carte (una, non numerata, contenente il titolo e una breve introduzione, le seguenti numerate da uno a 17). Le
carte sono rilegate insieme ad altre, contenenti altri tre racconti, all'interno di una copertina di cartoncino azzurro recante sulla prima un quadro sormontato dal titolo LE VERGINI DEL
SUD. Sotto l'illustrazione, l'indicazione 'di GIUSEPPE FAVA'. La carta utilizzata è molto sottile e ad alta trasparenza.
                     Si possono distinguere almeno cinque penne di colori diversi con cui l'autore interviene sul testo dattiloscritto e sul verso delle pagine per correggere e
integrare.</p>
               </objectDesc>
               <handDesc>
                  <handNote xml:id="FavaI">Scrittura autografa di Giuseppe Fava, penna di colore blu molto scuro, usata spesso per correzioni e integrazioni immediate del testo
base</handNote>
                  <handNote xml:id="FavaB" medium="blue_ink">Scrittura autografa di Giuseppe Fava, penna blu</handNote>
                  <handNote xml:id="FavaN" medium="black_ink">Scrittura autografa di Giuseppe Fava, penna nera</handNote>
                  <handNote xml:id="FavaR" medium="red_ink">Scrittura autografa di Giuseppe Fava, penna rossa</handNote>
                  <handNote xml:id="FavaV" medium="green_ink">Scrittura autografa di Giuseppe Fava, penna verde</handNote>
               </handDesc>
               <typeDesc>
                  <typeNote xml:id="Fava">Testo base dattiloscritto da Giuseppe Fava</typeNote>
               </typeDesc>
            </physDesc>
```

Figura 1. Dettaglio del `<teiHeader>` che presenta la descrizione del supporto fisico (`<physDesc>`).

Per evitare di appesantire il testo e renderlo di difficile lettura, non sono stati segnalati internamente al `<body>` gli interventi non latori di varianti. Nello specifico, l'autore interviene sistematicamente con la penna blu molto scuro per correggere refusi, per ripassare sopra parole impresse con un inchiostro troppo chiaro e per terminare parole rimaste incomplete in fine di rigo. Tutti i criteri adottati nell'allestire l'edizione sono stati segnalati nell'`<editorialDecl>`. Nello stesso luogo sono segnalate le correzioni di refusi e gli interventi la cui responsabilità è degli editori, come il ripristino quando necessario di punti fermi e la normalizzazione degli accenti.

Il frontespizio del racconto, costituito dal titolo e una breve premessa, è stato trascritto nel `<front>`.

Il testo, riprodotto nel `<body>`, è stato annotato utilizzando il tag `<add>` per le lezioni aggiunte e il tag `<del>` per le cassature. Qualora una cassatura e un'aggiunta siano simultanee (ad esempio nel caso di una parola scritta sopra una precedente) essi sono contenuti nel tag `<subst>`. Questi tre tag sono sempre completati con l'attributo `@hand` per indicare la penna utilizzata. Il tag `<add>` contiene, inoltre, l'attributo `@place` per segnalare la topografia della lezione, nei casi più frequenti "above" se nell'interlinea superiore, "inline" se in linea e "overwritten" se soprascritto alla cassatura.

In presenza di varianti adiafore, a `<add>` è stato aggiunto l'attributo `@type` completato dalla specificazione "adiafora".

Quando non è stato possibile leggere sotto la cassatura, si è utilizzato il tag `<unclear>` all'interno di `<del>`.

Laddove l'autore cambia idea e lascia parole incomplete è stato usato il tag `<supplied>` per integrare la parte mancante.

Nei casi di stratificazione correttoria particolarmente complessa, nei quali l'indicazione del colore della penna con cui l'autore è intervenuto è sembrata insufficiente, si è ricorso, oltre all'indicazione dell'attributo `@hand`, all'attributo `@seq` associato ad un numero. Con `@seq` si sono distinte le diverse fasi in cui l'autore torna su uno stesso passo, correggendo e integrando interventi precedentemente realizzati con la medesima penna o con una differente. Nell'esempio che segue (vd. Fig. 2) si sono distinte due penne (blu e nera) e cinque diverse fasi (vd. Fig. 3).

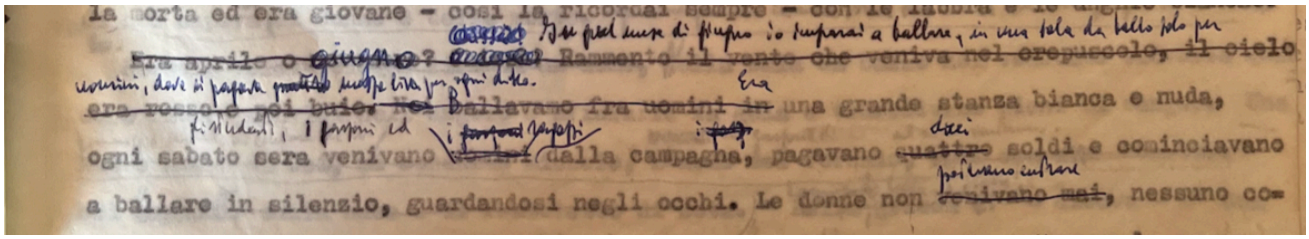


Figura 2. Esempio della stratificazione di interventi con penne di colori diversi

Figura 3. Esempio di codifica del testo con indicazione degli attributi `@hand` e `@seq`

Altre proposte di codifica

L'obiettivo auspicato per il progetto è di riuscire a fornire per ogni testo digitalizzato una codifica che ne rispecchi le peculiarità, incentrata di volta in volta sugli aspetti che si ritengono meritevoli di approfondimento.

Nel racconto *La ragazza che fu uccisa in luglio*, il narratore non fa mai riferimento a luoghi e anni specifici, usa espressioni come «piccolo paese sulla montagna» e «quel tempo».[8] A questa ambientazione indefinita sono, però, perfettamente sovrapponibili luoghi reali ed eventi storici legati alla biografia dell'autore: la storia si svolge infatti a Palazzolo Acreide nel luglio 1943, quando il paese natale dell'autore, allora diciassettenne, venne bombardato dagli Alleati nel corso delle operazioni militari legate allo sbarco in Sicilia. Nel `<teiHeader>`, all'interno del `<settingDesc>` sono presenti i tag `<place>` e `<date>`; per ogni luogo e data individuabile è stato creato un `@xml:id` e sono state fornite le informazioni relative agli avvenimenti storici correlati nel tag `<note>`.

La maggior parte dei personaggi non ha un nome, eppure in mezzo al coro degli abitanti del paese e dei soldati spiccano alcune figure: l'anonimo protagonista e narratore, la prostituta Carmela e la figlia Itria (la ragazza a cui allude il titolo) e altri caratteri che ricorrono nelle successive stesure del racconto, come il ragazzo che viene fucilato, il soldato inglese e il capitano. All'interno del `<particDesc>`, i tag `<person>` contengono un `@xml:id` e un tag `<note>` per ciascun personaggio significativo (vd. Fig. 4).

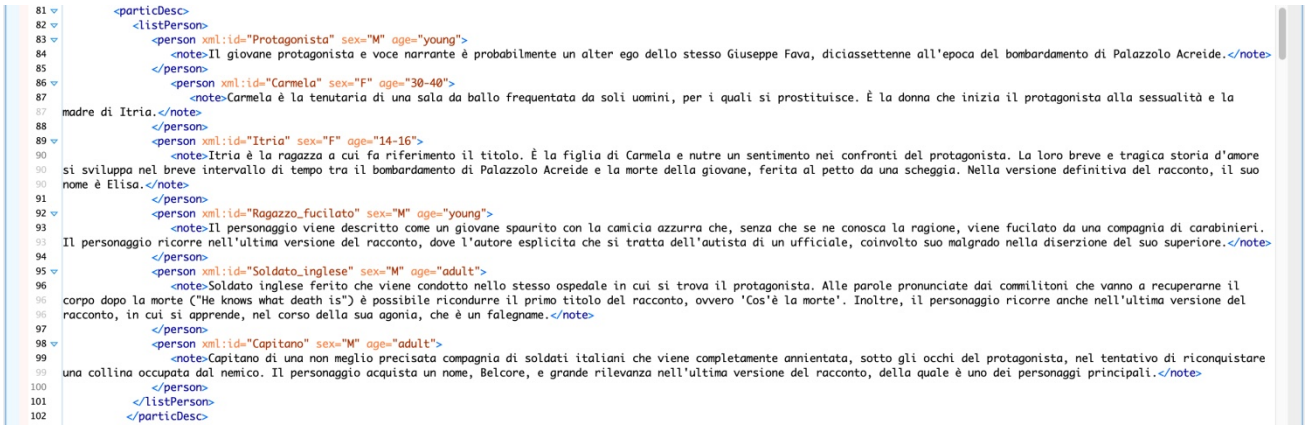


Figura 4. Dettaglio del `<teiHeader>` che presenta la descrizione dei personaggi significativi (`<particDesc>`)

5. CONCLUSIONI

L'Archivio di Giuseppe Fava è innegabilmente una risorsa di grande valore culturale che richiede adeguate misure di conservazione e promozione. Come si è potuto vedere, il progetto da noi elaborato mira a utilizzare gli strumenti del digitale per raggiungere questi due importanti obiettivi. Esso si presta inoltre a interessanti sviluppi grazie alla congiuntura favorevole tra possibilità offerte dall'umanistica digitale e caratteristiche intrinseche dell'opera di Fava. Questa, infatti, contraddistinta dall'eterogeneità delle forme praticate, dalla compattezza tematica e dal legame con i fatti storici e di

[8] L'unico riferimento all'anno in cui è ambientato il racconto si trova nella breve premessa presente nel frontespizio.

cronaca, si presta al raffronto e allo studio comparato delle sue parti; ciò si traduce nella possibilità di strutturare il complesso digitale tratto dall'Archivio come un vero e proprio *knowledge site*, un portale in cui le singole parti dialoghino tra loro attraverso, ad esempio, indici tematici e *timeline* e possano essere arricchite da materiale multimediale (anche dello stesso Fava) o collegamenti esterni per evidenziare l'importanza dell'opera del palazzolese nel contesto storico in cui è stata prodotta. Si avrebbe in questo modo uno strumento ricco e facilmente fruibile attraverso cui non solo tramandare la memoria di Fava, ma anche e soprattutto valorizzarne nella sua interezza la complessa figura intellettuale, troppo spesso schiacciata sulla sagoma bidimensionale del giornalista vittima di mafia.
Il progetto è in fase di interlocuzione con la Fondazione Giuseppe Fava, la quale si è mostrata favorevole all'idea e disponibile a definire i modi per la sua realizzazione.

BIBLIOGRAFIA

[1] Cannavò, Rosalba. *Pippo Fava: cronaca di un uomo libero*. Catania: CUECM, 1990.

[2] Di Mauro, Giuseppe Davide. «Giuseppe Fava, narratore». Tesi di laurea triennale, Università di Catania, 2021.

[3] Dolei, Giuseppe. *Il caso Fava tra poesia e verità*. Roma: Editoriale Artemide, 2010.

[4] Finocchiaro, Marzia, (a cura di). *La maestra e il diavolo*. Atti della giornata di studi dedicata a Giuseppe Fava. La Spezia: Agorà Edizioni, 2002.

[5] Giuffrida, Milena, Christian D'Agata, Laura Giurdanella, e Pietro Sichera. «Pirandello Nazionale: per un nuovo modello di edizione digitale, collaborativa e integrata». *AIUCD*, 2021, 207-215. http://amsacta.unibo.it/id/eprint/6712/.

[6] Italia, Paola. *Editing Duemila. Per una filologia dei testi digitali*. Roma: Salerno, 2020.

[7] Italia, Paola. «Filologia d'autore digitale». *Ecdotica* 16 (2019): 202-216.

[8] Mancinelli, Tiziana, e Elena Pierazzo. *Che cos'è un'edizione scientifica digitale*. Roma: Carocci, 2020.

[9] Mori, Giovanna. Giuseppe Fava. *La pittura come documento, racconto e denuncia*. Catania: Fondazione Giuseppe Fava, 2019.

[10] Randazzo, Pierlorenzo, (a cura di). *La passione del comprendere. Arte, politica e teatro di Giuseppe Fava*. Milano-Udine: Mimesis, 2023.

[11] Randazzo, Pierlorenzo. *La scena rivoluzionaria di Giuseppe Fava*. Palermo: Navarra Editore, 2023.

[12] Sahle, Patrick. «What is a Scholarly Digital Edition?» In *Digital Scholarly Editing: Theories and Practices*, (a cura di) Matthew James Driscoll e Elena Pierazzo, 19-39. Open Book Publishers, 2016. https://books.openedition.org/obp/3397?lang=it.

[13] Sichera, Antonio, e Antonio Di Silvestro. «"Pirandellonazionale" Una scommessa filologica ed ermeneutica». *Griseldaonline* 20, fasc. 2 (2021): 174-180. https://griseldaonline.unibo.it/article/view/12239/13460.

[14] Tomasi, Francesca. «Edizioni o archivi digitali? Knowledge sites e apporti disciplinari». In *Edizioni critiche digitali*, (a cura di) Paola Italia e Claudia Bonsi, 129-136. Roma: Sapienza Università Editrice, 2016. https://www.editricesapienza.it/sites/default/files/5369_Italia_Bonsi_EdizioniCriticheDigitali.pdf.

L'archivio digitale di una casa editrice: l'esempio del Saggiatore e della sua prima pubblicazione

Giada Di Pino
Università degli Studi di Catania, Italia - giada.dipino@phd.unict.it

ABSTRACT

Il presente contributo propone la descrizione di archivio digitale di una casa editrice, prendendo a modello il catalogo del Saggiatore. L'obiettivo è quello di creare un prototipo di archivio digitale integrato che possa contenere le notizie principali relative alla storia della casa editrice e alle figure storiche che l'hanno determinata e che hanno ruotato intorno a essa, insieme a tecnici, collaboratori e autori. Al tempo stesso, esso avrà la funzione di mettere in dialogo le pubblicazioni prodotte con la loro storia redazionale, con i processi decisionali, con i progetti editoriali e con le dinamiche di marketing e di produzione, per mezzo del materiale documentario e dei carteggi rinvenuti. Infatti, capire i processi decisionali e produttivi che hanno permesso alle opere di essere date alle stampe consente di ridisegnare i contorni del canone letterario odierno e della sua formazione. Inoltre, l'archivio si propone come un ulteriore strumento di studio e di ricerca delle opere letterarie e dei processi di formazione e di evoluzione che le hanno consegnate al pubblico di lettori.

PAROLE CHIAVE

Digital archives; publishing house; il Saggiatore; Thomas Mann; coding.

1. INTRODUZIONE

Parlare di storia della letteratura italiana, e in particolar modo della letteratura italiana contemporanea, significa parlare anche di storia dell'editoria. Le case editrici nel secondo Novecento hanno avuto l'importante ruolo non solo di assecondare, ma anche di indirizzare i gusti del pubblico, contribuendo in tal modo, anche in virtù di un sistematico e programmatico progetto culturale, alla creazione e al consolidamento del canone letterario per come lo conosciamo oggi.

Le Digital Humanities hanno prodotto i micro-universi delle edizioni digitali delle opere dei maggiori scrittori del Novecento e i loro archivi, quali, ad esempio, Pirandello Nazionale[1], l'archivio pascoliano (adesso offline), Manzoni.org[2] e Leggo Manzoni: Quaranta edizioni della Quarantana[3], le lettere di Vespasiano da Bisticci[4] e, per concludere con le piattaforme di edizione-archivio, il Bellini Digital Correspondance (BDC)[5], ma anche piattaforme per lo studio della filologia d'autore, come Philoeditor[6], o per la ricerca bibliografica, come Project Muse[7] e OPAC SBN[8], il servizio bibliotecario nazionale [15].

Un'ulteriore strada da percorrere potrebbe essere la creazione di archivi storici digitali delle case editrici, cioè di luoghi virtuali in cui la storia editoriale delle opere può affiancarsi all'edizione multimediale delle stesse, all'apparato documentario degli editori e al materiale redazionale; in cui, cioè, i testi possono essere visti sotto una luce diversa da quella comunemente gettata dallo studio del canone, e quindi dalla qualità letteraria di un'opera, ovvero quella del marketing editoriale, che mostra come anche la letteratura contemporanea sia in qualche modo frutto di un sistema capitalistico [1].

Il presente contributo vuole, dunque, proporre un modello possibile di archivio digitale di una casa editrice, prendendo come esempio di caso-studio quella fondata da Alberto Mondadori nel 1958: Il Saggiatore. Grazie alla sua storia, curata e periodicamente aggiornata da Alberto Cadioli [2], e all'ultima edizione del *Catalogo generale* [8], pubblicata in occasione dei sessant'anni dalla fondazione, è possibile avere un quadro dettagliato del significativo contributo che *Il Saggiatore*, con la sua politica editoriale apertamente votata alla diffusione e alla democratizzazione della cultura, ha apportato alla costituzione del canone letterario odierno, e, al tempo stesso, seguire nel tempo le dinamiche editoriali che hanno accompagnato l'ideazione delle collane e dei progetti redazionali all'interno della casa editrice. A partire quindi da tali

[1] https://www.pirandellonazionale.it/
[2] https://www.alessandromanzoni.org/
[3] https://www.projects.dharc.unibo.it/
[4] https://www.storiadigitale.it/
[5] https://bellinicorrespondence.cnr.it/
[6] https://www.filologiadautore.it
[7] https://www.muse.jhu.edu/
[8] https://www.opac.sbn.it/

linee guida, si può delineare un profilo dei testi pubblicati, analizzandoli caso per caso. In tal modo è possibile creare uno strumento di studio e di raccolta dati utile agli studiosi e ai ricercatori dell'ambito umanistico e non solo.

2. IL SAGGIATORE DI ALBERTO MONDADORI

Nel marzo del 1958 Alberto Mondadori scrive una lunga e dettagliata lettera a William Faulkner in cui annuncia il suo prossimo progetto di una nuova casa editrice, che avrebbe pubblicato il suo primo volume proprio nel corso di quello stesso anno. Il nome, evocativo e incisivo, è Il Sagittario. Questo è considerato l'atto di nascita ufficiale della casa editrice che oggi conosciamo con il nome del Saggiatore. Tuttavia, da alcuni carteggi con il padre, conservati oggi nell'archivio della Fondazione Arnoldo e Alberto Mondadori[9], si evince come già da almeno un anno l'idea avesse preso corpo, anche a causa del rapporto difficile tra padre e figlio. Alberto Mondadori, infatti, era già direttore editoriale (per l'esattezza, dal 1943) dell'omonima casa editrice al fianco del padre, ruolo che mantenne per tutta la vita [13].

Fin dalle prime lettere in cui ne dà notizia, il suo fondatore ha ben chiara la linea editoriale che deve mantenere la sua casa editrice: «La cultura, insomma, vista approfondita e interpretata attraverso la *storia*, la *critica*, i *testi*»[10] [12]. Il nome, invece, subisce una trasformazione. Il Sagittario, infatti, scelto da Alberto perché suo segno zodiacale, era anche il titolo di una fortunata collana nata pochi anni prima della Ceschina Editrice dedicata agli scrittori emergenti; così, viene scelto Il Saggiatore, che ha un doppio significato: da un lato è un'esplicita dedica all'opera di Galilei, simbolo per eccellenza di precisione scientifica e di libertà di pensiero, dall'altra è una citazione di Montaigne, che sosteneva che si legge e si scrive per "saggiarsi" e non per puro ozio [7]. Del Sagittario, tuttavia, è rimasta memoria nel logo per come lo conosciamo ancora oggi nella freccia che lo attraversa [4]. La prima collana pubblicata, la «Biblioteca delle Silerchie», la cui direzione era affidata a Giacomo Debenedetti [6], prende invece nome da via delle Silerchie, a Camaiore, dove sorgeva Villa Mondadori. La collana doveva contenere libri brevi, incisivi, poco noti e di autori già facenti parte del canone contemporaneo; il primo titolo, non a caso, è un testo di Thomas Mann, *Lettera sul matrimonio*. Seguono altre collane come «Uomo e Mito», «I Gabbiani», «La galleria del Minotauro», «Specchio del mondo» e, infine, «la Cultura», l'unica ad oggi sopravvissuta. Il Saggiatore, infatti, all'inizio degli anni Duemila ha compiuto una scelta editoriale tanto saggia quanto azzardata: si è posta come baluardo della cultura, appunto, chiudendo tutte le altre collane e pubblicando tutti i titoli nell'unica rimasta, sotto l'egida della «Cultura». Una cultura onnicomprensiva di saggistica, poesia, narrativa italiana e straniera, e avente come unico denominatore l'accessibilità; una cultura alla portata di tutti, contraddittoria e democratica come la società in cui viviamo.

3. L'ARCHIVIO

Il materiale documentario si compone principalmente di carteggi, non solo di Alberto Mondadori con gli autori, con Erich Linder, con i suoi collaboratori e con il personale della casa editrice, ma spesso anche con i suoi stessi familiari, con i familiari degli autori (rimandiamo all'esempio riportato nel paragrafo 6, in cui si analizza una lettera inviata da Alberto Mondadori a Katia Mann, la vedova dello scrittore) e con gli intellettuali di spicco del secondo Novecento. Tuttavia, si prevede anche la presenza di bozze, pareri di lettura, contratti editoriali e lavorativi, note di magazzino, notule di pagamento, conteggi, appunti, cartoline, telegrammi, agende e pagine di diario, oltre che le stesse opere pubblicate. Inoltre, nell'archivio saranno compresi metadati quali schede biografiche, paratesti, schede descrittive dei documenti stessi, approfondimenti, un ricco apparato fotografico ed eventuali contenuti multimediali, quali podcast e audiolibri.

L'archivio sarà suddiviso in tre sezioni: la storia della casa editrice, in cui sarà raccolto tutto il materiale documentario relativo alla nascita e all'evoluzione del Saggiatore; le collane e i testi. Il materiale documentario sarà dunque organizzato in tal senso: ciò che concerne la crescita e l'evoluzione nel tempo del Saggiatore, ciò che afferisce alle singole collane e alla loro formazione, e i testi, a loro volta raggruppati per collana e della cui organizzazione rimandiamo la descrizione al paragrafo 5, in cui proponiamo un esempio.

Tutto il materiale documentario, testuale, fotografico e sonoro contemplato per approfondire e indagare l'impatto culturale del Saggiatore nel panorama letterario, sebbene soggetto a eventuali espansioni o modifiche, sarà raccolto in un archivio digitale, che permetterà dunque un'organizzazione dei dati flessibile e scalabile e che possa contenere sia file in formato XML che file multimediali di varia natura. La maggior parte del materiale documentario, infatti, sarà archiviato come trascrizione in formato XML, così come anche i metadati esplicativi e di corredo. Tale formato, tuttavia, non è contemplato per i testi che non possono essere riprodotti integralmente, poiché ancora sotto la tutela del diritto d'autore. In questi casi, dunque, sono previsti gli incipit delle opere in dei file pdf realizzati con immagini ad alta risoluzione e digitalizzazione con

[9] https://www.fondazionemondadori.it/

[10] Lettera dattiloscritta di Alberto ad Arnoldo Mondadori. Zurigo, 29 gennaio 1958.

OCR, così come per gli articoli di approfondimento ed eventuale altro materiale d'apparato. Per quanto concerne invece il materiale fotografico, esso sarà digitalizzato in un formato ad alta risoluzione d'immagine.
Si prevede dunque un'interfaccia web sviluppata appositamente che dia la possibilità di ricercare tali dati, organizzati per documenti, per opere, per schede biografiche/autori, per luoghi. A livello di back-end, ogni pagina conterrà nel Tei Header i link di riferimento. È importante specificare, in tal senso, che l'archivio avrà una rispondenza per quanto possibile interna, dunque, ad esempio, i tag `<persName>` e `<title>` conterranno un attributo `@ref` con i link di rimando alle schede biografiche e delle opere già contenute in database.

4. IL SITO WEB IN OPEN ACCESS

Il prototipo dell'archivio digitale in open access si propone come uno strumento non solo di ricerca e di studio, ma anche di conservazione della memoria storica della casa editrice. Per tale motivo, esso, nella sua interfaccia web, è stato pensato come un'estensione del sito del Saggiatore[11], riproducendone la grafica, i colori bianco e rosso che caratterizzano anche le copertine delle pubblicazioni odierne, e la struttura, così da avere anche un alto grado di usabilità, pur essendo un ipertesto complesso e ricco di contenuti e link di collegamento interni (vd. Fig. 1). Si propone di seguito un esempio di struttura del portale web.

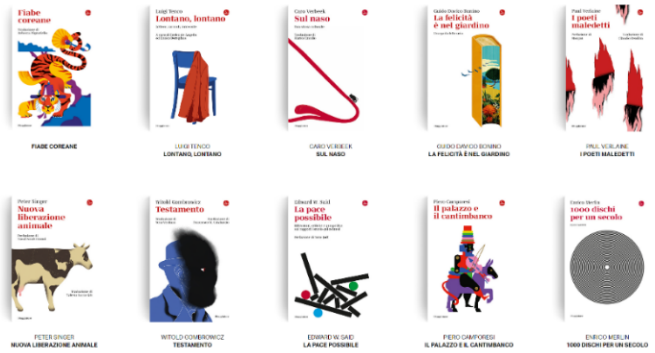


Figura 1. L'attuale disposizione del sito

Come anche la struttura interna dell'archivio digitale, anche l'interfaccia web rispecchierà la suddivisione in tre sezioni. Nella prima, "La casa editrice", sarà consultabile una breve storia del Saggiatore, dalla sua nascita fino ai giorni odierni, il materiale documentario, la linea editoriale, la descrizione e il progetto di marketing e culturale che sta dietro alla grafica delle odierne copertine, le schede biografiche, ordinate alfabeticamente, delle persone fisiche citate nello stesso archivio, strutturate in dati anagrafici e breve biografia; nel caso degli autori, saranno presenti anche una sezione sulle opere e un breve apparato critico e ogni scheda sarà infine corredata di una bibliografia.
La seconda sezione sarà dedicata alle collane, e a ciascuna di essa sarà associata una pagina di presentazione, cioè una breve descrizione delle pubblicazioni che conteneva e del progetto culturale che la sosteneva. Conterrà inoltre la "storia della collana", con la relativa descrizione della sua evoluzione dalla progettazione fino alla chiusura; i "documenti", con il materiale archivistico relativo; gli approfondimenti; la sezione "testi". Di questa terza e ultima sezione si rimanda la descrizione al paragrafo successivo. Infine, è prevista un'interfaccia di ricerca, che permetta all'utente del sito di effettuare una ricerca del materiale mirata e consapevole.

5. I TESTI: L'ESEMPIO LETTERA SUL MATRIMONIO DI THOMAS MANN

Dalla sezione "Testi" (vd. Fig. 1) si potrà accedere alle pagine dedicate alle opere. Come esempio da prototipo, si propone il primo testo pubblicato dalla casa editrice nell'ottobre del 1958: *Lettera sul matrimonio* di Thomas Mann [10]. L'opera fa parte, a sua volta, della prima collana inaugurata dal Saggiatore, «La Biblioteca delle Silerchie», di cui al paragrafo 2, ed è stata scelta da Alberto Mondadori proprio per la notorietà e la diffusione in Italia delle opere dello scrittore tedesco in quegli anni. Infatti, già la Arnoldo Mondadori Edizioni stava pubblicando un'opera omnia in più volumi la cui curatela era affidata alla nota germanista Lavinia Mazzucchetti. A lei si rivolse Alberto per trovare un testo adatto alla appena nata «La Biblioteca delle Silerchie», un'opera cioè che fosse breve e che non fosse stata ancora pubblicata nel territorio italiano. Lavinia Mazzucchetti consigliò *Lettera sul matrimonio*, che fu associato all'interno del volumetto al più noto *Brindisi a Katia*, un piccolo omaggio che lo scrittore dedicò alla moglie in occasione del suo settantesimo compleanno, ed entrambi furono tradotti da Italo Alighiero Chiusano, che allora stava muovendo i primi passi all'interno della grande editoria, e poi accolti nel volume dell'opera omnia sugli *Scritti minori* di Thomas Mann, che vide la pubblicazione ad appena un mese di distanza dal volumetto del Saggiatore [11].
Il testo sarà accessibile per mezzo di una pagina di presentazione in cui l'utente potrà immediatamente prendere visione delle informazioni principali: copertina, autore, titolo, una breve sinossi, curatore, traduttore e grafico, anno di pubblicazione, titolo originale ed eventuali altri dati nella parte centrale della pagina, mentre a destra sarà visibile un menù

[11] www.ilsaggiatore.it

laterale. Sarà possibile, inoltre, accedere a un pdf dell'intero libro o di una parte di esso, considerando anche lo stato dei diritti d'autore caso per caso, sfogliabile con un movimento d'immagine tridimensionale; i nomi di autore, curatore, traduttore, grafico o altri eventuali saranno a loro volta link di collegamento alle singole schede biografiche.
Per ogni opera è previsto l'accesso a ulteriore materiale d'archivio, per mezzo di un menù specifico. La sezione "Testo", in particolare, contiene la trascrizione dell'opera marcata secondo lo schema di codifica XML/TEI. La fase di codifica permette di attuare una riflessione critica sul testo e di formalizzare dunque il linguaggio di marcatura. Tuttavia, al fine di creare un sistema quanto più possibile circolare e rispondente a sé stesso, è necessario che il set di marcatori sia funzionale alle rispondenze interne dell'archivio. Lo strumento di visualizzazione prescelto sarà il software open source EVT1.3, che permette di confrontare il testo in trascrizione con la prima edizione del Saggiatore [14]. È importante specificare che questo contenuto non è previsto per tutti i testi della collana, ma solo per quelli che rispondono a specifiche condizioni, tra cui l'essere fuori dai diritti sia d'autore che di edizione e di traduzione, come nel caso dell'esempio specifico considerato, la lunghezza e la complessità del testo, la gestione della formalizzazione del linguaggio di marcatura.
Nella sezione "Edizione originale" è prevista, invece, in file in pdf, la prima edizione a stampa del testo, in lingua originale nel caso delle opere straniere; nel caso preso in esame, del volume antologico *Das Ehe-Buch*, a cura del conte Keyserling nel 1925 [9]. Saranno poi presenti le sezioni: "Documenti", per la cui descrizione rimandiamo al paragrafo 6; "Storia dell'edizione", in cui verrà brevemente delineata la storia redazionale del volume; "Edizioni", contenente una breve descrizione delle pubblicazioni, italiane ed estere, successive e precedenti, in cui l'opera appare; "Recensioni", dove sarà possibile raccogliere e catalogare le recensioni al testo reperite nella fase di ricerca; "Strumenti di promozione", in cui saranno inserite in un formato ad alta risoluzione, le fotografie delle pubblicità, perlopiù su rivista, attuate per la promozione del volume; "Lettura multimediale", che prevede il contenuto di file perlopiù audio, quali l'audiolibro dell'opera; "Approfondimenti", da strutturare caso per caso e in cui poter organizzare contenuti di varia natura, da podcast di approfondimento, appunto, a video ad articoli, eccetera; e, infine, "Dove trovarlo", nella forma di un nodo di collegamento alla pagina OPAC SBN corrispondente, come link esterno al sistema d'archivio.
Le diverse sezioni, soprattutto quelle relative alla nota storia dell'edizione, e le schede biografiche corrispondenti saranno corredate dall'apparato fotografico rinvenuto presso l'Archivio Lavinia Mazzucchetti, all'interno della Fondazione Arnoldo e Alberto Mondadori.

6. I DOCUMENTI. L'ESEMPIO DELLA LETTERA DI ALBERTO MONDADORI A KATIA MANN

La sezione "Documenti" è snodo fondamentale e di accesso, come si è potuto evincere dai paragrafi precedenti, dalle diverse sezioni ed è specifica per ciascuna di esse. Inoltre, ogni documento è rintracciabile anche per mezzo della barra di ricerca, che è esemplificativa di un'organizzazione dei documenti nel database per persone, luoghi, opere citate, mittente e destinatario, tipo di documento. L'apparato documentario non è composto, infatti, da soli carteggi: esso comprende anche bozze, contratti editoriali, appunti, note e dati di magazzino, e ciascun documento sarà codificato. Prendiamo come esempio di riferimento una lettera scritta da Alberto Mondadori a Katia Mann, la vedova dello scrittore, nel settembre del 1960.
La sezione "Documenti" relativa all'opera in esame contiene, come sarà per gran parte del materiale d'archivio, fonti documentarie provenienti dalla Fondazione Arnoldo e Alberto Mondadori, e in particolare sarà costituita da: documenti dell'Archivio Lavinia Mazzucchetti, soprattutto per il corredo fotografico, composto da numerose fotografie che ritraggono Alberto Mondadori, Lavinia Mazzucchetti e la famiglia Mann, ma anche Italo Alighiero Chiusano ed Ervino Pocar, dal carteggio tra Lavinia Mazzucchetti e Italo Alighiero Chiusano, e da alcune lettere che la germanista ha scambiato con l'Archivio e con il Centro Studi Thomas Mann; documenti provenienti dal fondo denominato "Il Saggiatore e altre società del Gruppo Il Saggiatore", costituiti prevalentemente dalle lettere di Alberto Mondadori a Lavinia Mazzucchetti, ad alcuni componenti della famiglia Mann e a Giuseppe Raimondi; i dati di magazzino dell'opera, provenienti invece dal fondo AME (Arnoldo Mondadori Editore). Nel Tei Header di marcatura è prevista la codifica delle informazioni riguardati la fonte d'archivio del documento in esame, con il tag `<objectDesc>`, contenente la tipologia di documento e il fondo da cui proviene, nella sottocategoria `<repository>`.
Ricordiamo che essendo il presente progetto in forma di prototipo, ed esso stesso ancora in forma di sistematizzazione, ed essendo inoltre soggetto alla revisione del modello da parte dell'equipe tecnica del Saggiatore, si profila qui un possibile forma di struttura di back-end e di front-end non definitivi né esaustivi.

L'intero documento preso in esame sarà catalogato nel database insieme alla sua trascrizione, con codifica XML/TEI effettuata in parte con Oxygen e in parte con Leaf-Writer online, software che permette di trattare l'XML del testo con una selezione di marcatori automatici. Per la codifica sono stati usati i tag base del documento "lettera", inseriti all'interno del tagset `<correspDesc>`, in cui, oltre ai comuni `<text>`, `<body>`, `<div>`, corredato da attributo `@type=letter`, `<head>` e la suddivisione del corpo della lettera in `<p>`, si possono notare i tag `<opener>` per la formula di apertura, `<dateline>` per l'impostazione di data e luogo, `<salute>` per la formula di apertura; dopo il corpo testo, invece, si è usato `<closer>` per la formula di chiusura, al cui interno si evidenziano `<salute>` per i saluti conclusivi e `<signed>` per la firma del mittente; a seguire, `<epigraph>` contiene il nome e l'indirizzo del destinatario (vd. Fig. 2) [5]. Per la marcatura, come si accennava, Leaf-Writer permette un sistema di visualizzazione dei tag facile e intuitivo (vd. Fig. 3). Nel caso specifico sono stati utilizzati i marcatori `<persname>` per indicare le persone fisiche citate, `<placename>` per i luoghi, `<date>` con attributo `@when` per le indicazioni temporali esplicite. Si segnala inoltre l'uso di `<orgname>` per indicare in questo caso le case editrici menzionate nella lettera, `<title>` per il riferimento alla collana e all'opera. Inoltre, sono state evidenziate le parole chiave, marcate come `<term>`, "lettera" e "edizione", che fanno riferimento, rispettivamente, allo scambio di lettere tra l'editore e la moglie dello scrittore e alle edizioni dell'opera di Mann presa a modello. Man mano che l'archivio sarà costituito, gli attributi `@ref` potranno contenere i link interni all'archivio stesso.

```
2115  ]]>
2116        </rdf:Description>
2117     </rdf:RDF></xenoData></teiHeader>
2118     <text>
2119        <body>
2120           <div type="letter">
2121              <head>Lettera di <persName key="Mondadori, Alberto, 1914-1976" ref="http://viaf.org/
                  viaf/9973589/">Alberto Mondadori</persName> a <persName key="Katia Mann"
                  ref="http://www.wikidata.org/entity/Q214999">Katia Mann</persName></head>
2122              <opener>
2123                 <note type="setting">
2124                    <p><placeName key="Fondazione Arnoldo e Alberto Mondadori" ref="http://www.
                        wikidata.org/entity/Q81173243">Fondazione Arnoldo e Alberto Mondadori</
                        placeName>.</p>
2125                    <p>Fascicolo <persName key="Katia Mann" ref="http://www.wikidata.org/entity/
                        Q214999">Mann Katia</persName>, sezione <persName key="Mondadori,
                        Alberto, 1914-1976" ref="http://viaf.org/viaf/9973589/">Alberto
                        Mondadori</persName>, carteggio</p>
2126                 </note>
2127                 <dateline>
2128                    <placeName key="Milan" ref="http://www.wikidata.org/entity/Q490">Milano</
                        placeName>, <date when="1960-09-29">29 settembre 1960</date>
2129                 </dateline>
2130                 <salute>Gentile e cara signora,</salute>
2131              </opener>
2132              <p>faccio seguito al mio telegramma di <date when="1960-09-28">ieri</date> per
                  scusarmi ancora una volta del ritardo col ritardo col quale rispondo alla Sua
                  lettera<seg type="keyword"><term><term><term/></term></term></seg> del <date
                  when="1960-09-07">7 settembre</date>. Nel frattempo però dovrebbe esserLe
                  giunta una lettera<seg type="keyword"><term><term><term/></term></term></seg>
                  del mio assistente, <persName key="Lettieri, Mario" ref="http://viaf.org/viaf/
```

Figura 2. Codifica in XML/TEI della lettera di Alberto Mondadori a Katia Mann

L'immagine del documento, visibile accanto al testo marcato, svolgerà la funzione di nodo di accesso all'edizione diplomatica del documento affiancata dalla scansione in alta risoluzione, al fine di garantire anche l'analisi dal punto di vista filologico. Nella maggior parte dei casi si tratta di testi dattiloscritti o manoscritti, e tale caratteristica sarà indicata nel Tei Header con il tag `<physDesc>` corredato dall'attributo `@type`. Inoltre, verranno utilizzati i tag `<del>` e `<add>` nel caso di eventuali correzioni apportate dall'autore, contenuti, qualora si tratti di sostituzioni, all'interno del tag `<mod>` e corredati di attributi quali `@place` e `@type`. La visualizzazione, come quella del testo dell'opera, sarà realizzata con EVT1.3.

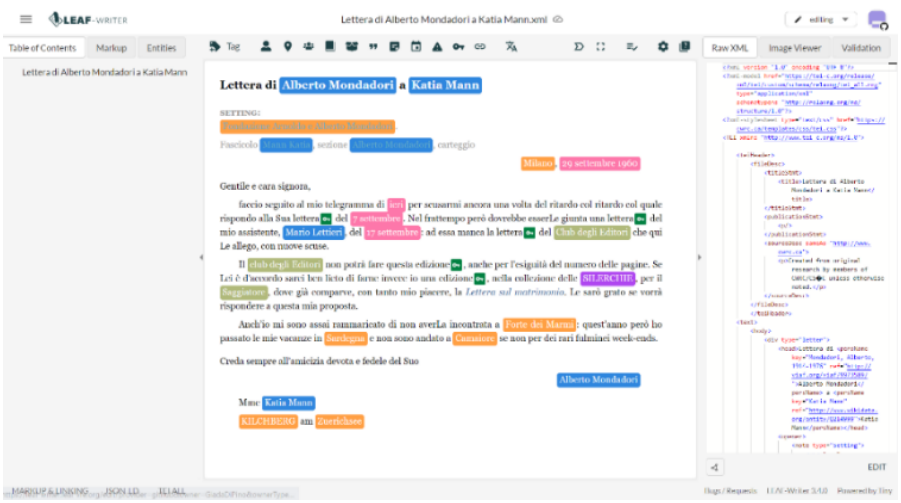


Figura 3. Esempio di visualizzazione di codifica con Leaf-Writer

La trascrizione di cui sopra, infine, sarà inserita all'interno di una schedatura di catalogazione al fine di mettere in evidenza le caratteristiche del documento. Nel caso preso in esame, ad esempio, saranno esplicitati mittente, destinatario, data, luogo, lingua, contenuto, regesto, edizioni, descrizione della carta, opere citate, persone citate, luoghi citati, che rimandano ai rispettivi luoghi del database non relazionale.

7. CONCLUSIONI

In conclusione, il lavoro che prospetta il presente progetto si sviluppa su tre direttrici: lo studio della storia della casa editrice, lo studio delle collane e lo studio delle pubblicazioni. Per tutte e tre le zone di ricerca è necessario distinguere una fase di ricerca e di studio delle fonti e una fase di stesura e produzione del materiale, oltre a una fase di digitalizzazione dello stesso. Dalla descrizione di cui ai paragrafi precedenti, si evince anche come il materiale sia potenzialmente in continuo dialogo e rispondenza, dunque il lavoro deve essere svolto in maniera quanto più organica e organizzata possibile. Per il prototipo presentato si è iniziato dalle origini, seguendo una linea cronologica e delimitando la ricerca a una collana; tuttavia, il potenziale di espansione di un archivio di siffatta maniera potrebbe condurre a lavorare sull'intero archivio della casa editrice, sulle opere letterarie come su quelle scientifiche e di saggistica, fino a giungere alle pubblicazioni più recenti, dando prova dunque di studi filologici basati sui testi dell'era digitale, e che dunque si configura già tale fin dalla nascita del testo letterario di cui si occupa [3]. Nonostante le prospettive possano risultare sconfortanti, per via proprio dell'entità delle possibili espansioni e della grande quantità di dati da elaborare, la sostenibilità di un tale progetto potrebbe essere garantita proprio dalla presenza del patrocinio della casa editrice, organo privato che tuttavia tende a mantenere un'immagine pubblica improntata al principio della serietà scientifica, dell'impegno culturale e della promozione del sapere.

Un archivio digitale come questo risulterebbe dunque essere, infine, uno strumento dalle molteplici funzioni e per più destinatari: per la casa editrice, in primo luogo, che avrà modo di rinverdire e riutilizzare la sua memoria storica, di approfondirla e di renderla fruibile e accessibile; per i ricercatori e gli studiosi, non solo delle discipline umanistiche, ma anche di quelle scientifiche, che potranno avere accesso a una rete di dati e di informazioni utili sia per indagare le dinamiche editoriali alla base della costituzione del canone letterario e bibliografico, sia per costruire bibliografie rispondenti alle esigenze delle ricerche; per i lettori del Saggiatore, infine, che si caratterizzano principalmente per essere un pubblico fidelizzato e rispondente a target del lettore forte e informato.

8. RINGRAZIAMENTI

Il progetto è finanziato dall'Unione Europea – Next Generation EU, ed è realizzato grazie alla collaborazione tra l'Università di Catania e Il Saggiatore Si ringrazia inoltre la Fondazione Arnoldo e Alberto Mondadori, che ha concesso l'accesso agli archivi.

BIBLIOGRAFIA

[1] Bollo, Alessandro. *Il marketing della cultura*. Milano: Carocci Editore, 2019.

[2] Cadioli, Alberto. *Sono un esploratore, mi piace viaggiare nel tempo. Breve storia del Saggiatore dal 1958 a oggi*. Milano: il Saggiatore, 1993.

[3] Carbé, Emmanuela. *Digitale d'autore. Macchine, archivi, letteratura.* Firenze: Siena Firenze University Press, 2023.

[4] Cavalli, Arianna, e Giacomo Papi. *Cose spiegate bene a proposito di libri*. Milano: Iperborea, 2021.

[5] Ciotti, Fabio, (a cura di). *Digital Humanities. Metodi, strumenti, saperi*. Roma: Carocci, 2023.

[6] Debenedetti, Giacomo. *Preludi. Le note editoriali alla «Biblioteca delle Silerchie».* Palermo: Sellerio, 2012.

[7] *Il Saggiatore 1958-2018. Catalogo generale*. Milano: Il Saggiatore, 2018.

[8] *Il Saggiatore. Catalogo n. 2 primavera-estate 1959*. Milano: Il Saggiatore, 1959.

[9] Keyserling, Hermann. *Das Ehe-Buch.* Celle: Niels Kampmann Verlag, 1925.

[10] Mann, Thomas. *Lettera sul matrimonio*. Milano: Il Saggiatore, 1958.

[11] Mann, Thomas. *Scritti minori. Tutte le opere*. Milano: Mondadori, 1958.

[12] Mondadori, Alberto. *Lettere di una vita. 1922-1975.* Milano: Fondazione Arnoldo e Alberto Mondadori – Arnoldo Mondadori Editore, 1996.

[13] Palermitano, Andrea. *Storia del Saggiatore. I primi sessant'anni.* Milano: Il Saggiatore, 2018.

[14] Sahle, Patrick. *«What is a scholary digital edition?» In Digital scholarly editing: Theories and practices.* Cambridge: OpenBook Publishers, 2016. http://books.openedition.org/obp/3397.

[15] Stella, Francesco. *Testi letterari e analisi digitale.* Milano: Carocci, 2018.

Marcare la poesia del Novecento: uno studio per *Ossi di seppia*

Chiara Cauzzi[1], Martina Corti[2], Anna Guadagnoli[3], Maria Grazia Schiaroli[4]
[1] Università della Svizzera italiana, Istituto di studi italiani, Biblioteca universitaria Lugano, Svizzera - chiara.cauzzi@usi.ch
[2] Università degli studi di Siena, Italia - martina.corti@student.unisi.it
[3] Università degli studi di Siena, Italia - anna.guadagnoli@gmail.com
[4] Università degli studi di Siena, Italia - m.schiaroli@student.unisi.it

ABSTRACT
Il contributo esplora le potenzialità e i limiti della codifica XML/TEI per la rappresentazione delle varianti d'autore nella poesia del Novecento, proponendo come caso di studio *Ossi di seppia* di Eugenio Montale, una raccolta profondamente legata all'immaginario del Mediterraneo. Si focalizza in particolare sulla terza sezione eponima, la più ricca e complessa dell'opera, sulla scorta dell'edizione critica di Bettarini-Contini. Il contributo formula ipotesi di lavoro per rappresentare le varianti usando una combinazione di moduli TEI adattati alle esigenze della filologia d'autore. Si evidenziano così la validità e l'economicità della codifica XML/TEI per la poesia contemporanea, non senza alcune criticità per le varianti d'autore più complesse.

PAROLE CHIAVE
Edizione digitale; XML/TEI; poesia; filologia d'autore; Montale.

1. INTRODUZIONE
Nell'ultimo decennio, la rapida diffusione della filologia digitale ha favorito l'applicazione di metodi e strumenti informatici alla critica del testo e stimolato la pubblicazione di edizioni scientifiche nativamente digitali. Lungi dallo scopo di voler sostituire l'edizione a stampa, quella digitale, servendosi del mezzo elettronico, ha implementato le potenzialità della tradizionale attività esegetica su carta rendendo, in primo luogo, la consultazione più ricca e dinamica [2, 3].
In questo contesto, le sfide e le criticità poste dalla filologia digitale, sia della copia sia d'autore, sono state accolte dallo standard di codifica XML/TEI che questo contributo adotta per rappresentare le varianti di uno degli autori più significativi del Novecento [6]. Se, infatti, il panorama delle edizioni nativamente digitali è ricco di casi di studio appartenenti alla filologia della copia risalenti al Medioevo, meno pubblicazioni sono state dedicate alle edizioni critiche di filologia d'autore del Novecento [7]. Il contributo muove dall'obiettivo di applicare le potenzialità di XML/TEI in questo ambito ancora poco esplorato, partecipando alla riflessione intorno alla filologia d'autore e alla critica delle varianti del XX secolo.
La raccolta *Ossi di seppia* di Eugenio Montale [4] rappresenta un significativo caso di studio per la definizione del modello più appropriato di codifica XML/TEI che è stato applicato su alcuni dei componimenti più emblematici della terza sezione della raccolta dal titolo eponimo. La codifica è condotta sull'edizione critica Bettarini-Contini (1980) perché corredata da un rigoroso apparato critico che tiene conto delle numerose varianti d'autore, la cui rappresentazione digitale si colloca all'interno della più ampia riflessione su un modello di codifica per la poesia contemporanea [5].
La soluzione adottata si caratterizza per la sua economicità: il modulo *Verse*, che risolve le criticità legate alla divisione dei testi in versi, viene combinato con *Manuscript Description*, *Critical Apparatus* e *Representation of primary sources*, pensati sulla base delle necessità della filologia d'autore. Il primo raccoglie la bibliografia e aiuta a ricostruire la vicenda compositiva dell'opera, il secondo permette di gestire le voci di apparato, manoscritte o a stampa, mentre il terzo consente la rappresentazione delle modifiche autografe nei diversi testimoni.
Il risultato dell'operazione, come mostra il prototipo di codifica, è un'edizione critica che nulla toglie all'esattezza del modello cartaceo, piuttosto aggiunge alcune opportunità di visualizzazione e comprensione del lavoro stratificato dell'autore.

2. QUESTIONI DI MARCATURA
Lo standard di codifica XML/TEI rappresenta da tempo il punto di riferimento nell'ambito delle edizioni digitali: l'enorme varietà di tag a disposizione e la sua flessibilità lo rendono uno strumento versatile e in grado di adattarsi a testi estremamente diversi tra loro [1]. È curioso quindi quanto ancora sia stata poco esplorata la marcatura di poesia, specialmente della poesia italiana contemporanea. Ancora più complessa è la questione della marcatura delle edizioni critiche: non esiste, infatti, un modulo di XML/TEI dedicato alla filologia d'autore, per cui sorge il problema di come mostrare la storia di un testo tramite le sue varianti [9: 218-222]. Il progetto quindi di marcare un estratto dell'opera

montaliana si è aperto con la ricerca di un esempio di marcatura che tenesse conto sia della versificazione che della presenza di varianti.
La ricerca di un modello si è concentrata più sull'aspetto della filologia d'autore che su quello della poesia novecentesca: la presenza del modulo 6 delle linee guida XML/TEI, *Verse*[1], dedicato alla poesia, rende più chiare le possibili modalità di marcatura del verso. Per la filologia d'autore, invece, si è cercato di rendere fedelmente l'edizione critica a stampa attraverso i tag più comunemente adottati per marcare le varianti manoscritte. Imprescindibile in un'attività di questa natura è stato il lavoro di Paola Italia [2, 3], come è stato indispensabile il confronto con alcuni esempi di marcatura sia poetica che filologica: un esempio, è stato quello proposto durante il workshop *Codificare (al)l'Infinito*, tenutosi durante il convegno AIUCD 2023, ospitato a Siena, il cui modello di codifica dell'edizione critica dei *Canti* leopardiani è stato utile a risolvere alcuni dei problemi della filologia d'autore digitalizzata [8: 421-423].
La struttura della marcatura è stata organizzata per inserire i testi nel contesto della raccolta. Per questo motivo è stato usato un `<text>` per l'intera opera. Le poesie sono state interpretate come ulteriori `<text>`, uniti nelle diverse sezioni della raccolta tramite `<group>`. A questo punto si entra nel cuore del testo: ciascuna poesia è inserita in un tag `<lg>`, mentre i versi sono contenuti in un tag `<l>`, fornito di numerazione, indicazione metrica e id univoco e progressivo; se sono presenti delle strofe, i versi sono stati raccolti in un ulteriore tag `<lg>`.
Si è poi riflettuto su alcuni metodi per rappresentare nel linguaggio XML/TEI la varietà di casi di filologia d'autore offerti dall'apparato critico di Bettarini e Contini, insieme alle varianti del testo [5]. Una prima componente da considerare è quella dei testimoni. Nel `<teiHeader>` è stato di fondamentale importanza l'elemento del `<SourceDesc>`, grazie al quale è stato possibile creare quattro categorie di testi citati nell'edizione: una `<listBibl>` dedicata agli studi citati nell'apparato e una per le edizioni critiche; una `<listWit>` per i testimoni manoscritti, corredati dalle informazioni di conservazione e, quando disponibile, della descrizione fisica, e una per quelli a stampa. La tradizione del testo è accuratamente ricostruita nell'edizione dei due filologi: nel progetto di codifica se ne dà conto per ciascuna poesia in una nota dedicata, identificata da un `@type:text-history` e un id. Più complessa la questione delle varianti: si è fatto uso di alcuni elementi tipici del modulo 12 XML/TEI, *Critical Apparatus*[2]; in presenza di varianti, la voce di apparato è codificata tramite il tag `<app>`, all'interno del quale il testo definitivo si trova in `<lem>`, mentre le varianti in altrettanti `<rdg>`.
Il sistema può bastare nei casi in cui le varianti sono soltanto a stampa o per casi molto semplici di varianti manoscritte, non è tuttavia in grado di rappresentare tutte le complessità possibili: `<lem>` e `<rdg>` non sono infatti sufficienti per riportare la complessità del lavoro di scrittura che la filologia d'autore cerca di ricostruire. Un elemento distintivo della storia compositiva di una raccolta poetica è che ciascuna poesia può avere una vicenda specifica e fasi compositive diversificate, anche all'interno dello stesso testimone. Per codificare questo tipo di eventi testuali si è quindi fatto ricorso al modulo 10 XML/TEI, *Manuscript Description*[3]. Nel `<profileDesc>` del `<teiHeader>` è stata inserita la voce `<creation>`, in cui sono riportati i diversi `<change>` delle poesie, ognuno corredato da un id, dall'indicazione del testimone in cui appare e della tipologia di modifica (cancellatura, aggiunta, ecc.) e, quando disponibile, dalla datazione. Integrando i `<rdg>` con `@change` è possibile quindi associare a ciascuna modifica la fase compositiva specifica (vd. Fig. 1).

```
<change xml:id="meriggiare-ms1" type="alternative" source="#Deb">redazione
   manoscritta con varianti, provvista di titolo, interamente biffata mediante tratti
   obliqui incrociati, senza data, contenuta nel <rs type="witness" ref="#Deb"
      >fascicoletto</rs> inviato a <persName ref="#Debenedetti">Giacomo
      Debenedetti</persName>; in calce una Nota dell'<rs type="person" ref="#Montale"
      >autore</rs> contenente una variante per il <rs type="verse" ref="meriggiare-8"
      >v. 8</rs></change>
```

Figura 1

Per dare ulteriore esattezza alla marcatura, si è lavorato sull'inserimento nella codifica delle modifiche manoscritte fatte dall'autore, come cancellature immediate, sostituzioni o aggiunte a margine. Per questo è stato utilizzato `<mod>`, tag del modulo 11 XML/TEI, *Representation of Primary Sources*[4], insieme a `<del>` e `<add>` (vd. Figg. 2 e 3).
Questi sono stati gli elementi centrali nel tentativo di creare un modello di marcatura per un'edizione critica digitale degli *Ossi di seppia*. Oltre ai tag qui citati, è stato fatto un approfondito lavoro di codifica di ogni elemento delle note critiche:

[1] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/VE.html
[2] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html
[3] https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html
[4] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html

dalle varie stampe agli articoli di commento, dalle persone destinatarie dei manoscritti a amici e studiosi di Montale, ogni titolo e nome è andato a popolare la <listBibl> e la <listPerson> inserita nel <teiHeader>.
Il tentativo di integrare edizione ed apparato critico e filologia d'autore risulta possibile con la marcatura XML/TEI. Si è dunque tentato di visualizzare il risultato tramite il software EVT.

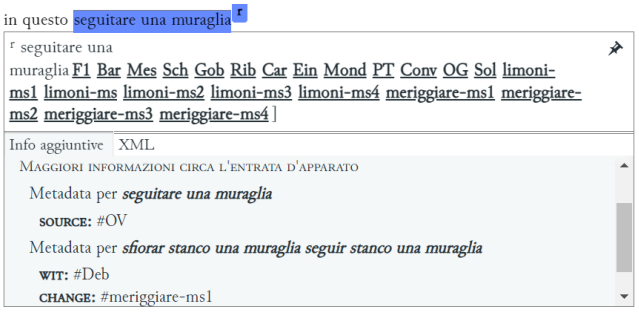
```
<app>
   <lem source="#OV" xml:id="limoni-lem1"><l n="1" xml:id="limoni-1"
         >Ascoltami, i poeti laureati</l>
      <l n="2" xml:id="limoni-2">si muovono soltanto fra le piante</l>
      <l n="3" xml:id="limoni-3">dai nomi poco usati<app source="#OV">
            <lem source="#OV" xml:id="limoni-lem2">:</lem>
            <rdg wit="#Rib #Car #Ein1">;<witDetail wit="#Rib #Car #Ein1"
                  >probabile svista tipografica</witDetail></rdg>
         </app> bossi ligustri o acanti.</l></lem>
   <rdg wit="#F1" change="#limoni-ms1"><l n="1" corresp="#limoni-1"
         >Ascolta,</l><l n="2" corresp="#limoni-1 #limoni-2">i poeti
         laureati si muovono soltanto</l><l n="3"
         corresp="#limoni-2 #limoni-3">tra <mod instant="true"><del>gli
               alberi</del><add>le piante</add></mod> dai nomi poco
         usati:</l><l n="4" corresp="#limoni-3">i ligustri, i bossi, o gli
         acanti...</l></rdg>
   <rdg wit="#Deb" change="#limoni-ms2"><l n="1" corresp="#limoni-1"
         >Ascolta,</l><l n="2" corresp="#limoni-1 #limoni-2">i poeti
         laureati si muovono soltanto</l><l n="3"
         corresp="#limoni-2 #limoni-3">fra le piante dai nomi poco
         usati:</l><l n="4" corresp="#limoni-3">i ligustri, i bossi, o gli
         acanti...</l></rdg>
   <rdg wit="#Bar #Mes" change="#limoni-ms3 #limoni-ms4"><l n="1"
         corresp="#limoni-1">Ascolta,</l><l n="2"
         corresp="#limoni-1 #limoni-2">i poeti laureati si muovono
         soltanto</l><l n="3" corresp="#limoni-2 #limoni-3">fra le piante
         dai nomi poco usati:</l><l n="4" corresp="#limoni-3">i ligustri,
         bossi, o gli acanti...</l></rdg>
</app>
```

```
<head type="title"><app>
      <lem source="#OV"/>
      <rdg wit="#Deb" change="#meriggiare-ms1">Tra gli orti</rdg>
      <rdg wit="#Sch" change="#meriggiare-ms2">Rottami</rdg>
   </app></head>
<div type="text">
   <lg xml:id="meriggiare-a">
      <l n="1" xml:id="meriggiare-1">Meriggiare pallido e assorto</l>
      <l n="2" xml:id="meriggiare-2">presso un rovente muro d'orto<app>
            <lem source="#OV" xml:id="meriggiare-lem1">,</lem>
            <rdg wit="#Deb" change="#meriggiare-ms1">;</rdg>
         </app>,</l>
      <l n="3" xml:id="meriggiare-3">ascoltare <app>
            <lem source="#OV" xml:id="meriggiare-lem2">tra i pruni e gli sterpi</lem>
            <rdg wit="#Deb" change="#meriggiare-ms1">tra pruni e sterpi</rdg>
         </app>
      </l>
      <l n="4" xml:id="meriggiare-4">schiocchi di merli, <app>
            <lem source="#OV" xml:id="meriggiare-lem3">frusci</lem>
            <rdg wit="#Sch #Bar #Mes"
               change="#meriggiare-ms2 #meriggiare-ms3 #meriggiare-ms4">sfrusci</rdg>
         </app> di serpi.</l>
   </lg>
```

Figure 2 e 3

3. VISUALIZZAZIONE IN EVT

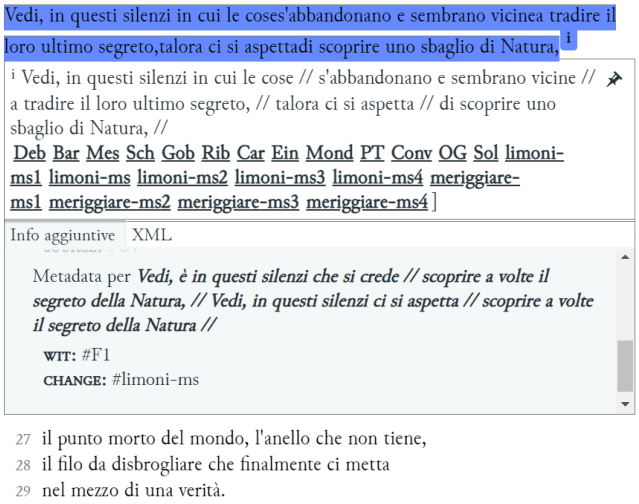
Con la codifica dei testi tratti dagli *Ossi di seppia* sono state rilevate alcune questioni relative alla visualizzazione che potrebbero, a nostro parere, essere di spunto per affinare ulteriormente la resa grafica di una futura versione di EVT.
La questione di più immediata evidenza è la mancata resa della divisione del testo in strofe. Nonostante l'uso del tag apposito, queste unità strutturali fondamentali non vengono adeguatamente separate l'una dall'altra. Anche la sostituzione di <lg> con l'elemento <milestone> dotato di @unit e valore 'stanza', soluzione adottata anche nella codifica del teatro in versi, non produce l'esito sperato.
In linea teorica, la scrittura di una regola ad hoc nel file .css potrebbe risolvere il problema, ma nell'esplorare codifiche alternative si è riscontrata un'incompatibilità tra <l> e <lb>: la compresenza degli elementi fa venire meno la divisione in versi. Un espediente efficace potrebbe essere la rinuncia ad <lb> in favore di <p>, nondimeno esso implicherebbe una perdita in termini di accuratezza, trattandosi pur sempre di un elemento semanticamente non appropriato.
Altra questione prioritaria è quella dell'apparato critico. Essendo stati presi in esame testi a stampa molto diffusi, per i <lem> si è preferito utilizzare l'attributo @source piuttosto che @wit. Tuttavia, laddove non venga specificato il testimone di una lezione, EVT sembra attribuirla a tutte le @wit disponibili; inoltre, i <rdg> figurano soltanto tra le informazioni aggiuntive, mentre eventuali modifiche manoscritte marcate con <mod> non vengono visualizzate correttamente, riportando testi cancellati e aggiunti senza alcuna distinzione (vd. Figg. 4 e 5).




```
<l n="16" xml:id="meriggiare-16">in questo <app>
      <lem source="#OV" xml:id="meriggiare-lem8">seguitare una muraglia</lem>
      <rdg wit="#Deb" change="#meriggiare-ms1"><mod><del>sfiorar stanco una
            muraglia</del><add>seguir stanco una muraglia</add></mod></rdg>
   </app></l>
```

Figure 4 e 5

L'indicazione delle varie modifiche nella sezione dedicata alle informazioni aggiuntive, renderebbe l'apparato fortemente informativo. Vi compaiono già, in effetti, i `<change>`, per i quali si renderebbe necessaria solamente una presentazione più ordinata. Nello stesso luogo si potrebbe dar conto anche delle aggiunte e cancellazioni, già gestite correttamente da EVT nelle edizioni diplomatiche.
Si vuole portare l'attenzione anche su un aspetto relativo all'elemento `<app>`. Qualora uno o più versi siano contenuti nella loro interezza in una voce di apparato, EVT restituisce un unico blocco testuale privo di numerazione e di distinzione tra la parola conclusiva di un verso e quella iniziale del successivo (vd. Figg. 6 e 7). Al fine di rendere il testo critico corretto e fruibile, sarebbe opportuno trovare una soluzione a questa criticità.




```
<app>
   <lem source="#OV" xml:id="limoni-lem9"><l n="22" xml:id="limoni-22">Vedi,
         in questi silenzi in cui le cose</l>
      <l n="23" xml:id="limoni-23">s'abbandonano e sembrano vicine</l>
      <l n="24" xml:id="limoni-24">a tradire il loro ultimo segreto,</l>
      <l n="25" xml:id="limoni-25">talora ci si aspetta</l>
      <l n="26" xml:id="limoni-26">di scoprire uno sbaglio di
      Natura,</l></lem>
   <rdg wit="#F1" change="#limoni-ms"><mod><del><l n="22"
               corresp="#limoni-22">Vedi, è in questi silenzi che si
               crede</l>
            <l n="23" corresp="#limoni-26">scoprire a volte il segreto della
               Natura,</l></del><add><l n="22" corresp="#limoni-22">Vedi, in
               questi silenzi ci si aspetta</l><l n="23"
               corresp="#limoni-26">scoprire a volte il segreto della
               Natura</l></add></mod></rdg>
</app>
<l n="27" xml:id="limoni-27">il punto morto del mondo, l'anello che non
   tiene,</l>
<l n="28" xml:id="limoni-28">il filo da disbrogliare che finalmente ci
   metta</l>
<l n="29" xml:id="limoni-29">nel mezzo di una verità.</l>
```

Figure 6 e 7

Infine, essendo oggetto di codifica specificamente l'edizione Bettarini-Contini, è essenziale poter leggere le note dei due studiosi, marcate avendo cura di inserire tutti i riferimenti bibliografici necessari e funzionali alla ricostruzione della storia dei testi. La strategia utilizzata a questo scopo è stata la sostituzione del tag `<note>` con un `<div>`, provvisto di `@type` specifico, alla fine di ciascun testo poetico.
La terza versione di EVT si propone di implementare nuove funzionalità legate alla risoluzione delle problematiche poste dalla filologia d'autore, presumibilmente, dunque, gran parte di queste verranno affrontate. Se ne attende la prossima uscita per poter effettuare nuove sperimentazioni sui testi scelti.

BIBLIOGRAFIA

[1] D'Agostino, Giulia, Giulia Fabbris, e Roberto Rosselli Del Turco. «Workshop sulle edizioni digitali: preparazione con codifica XML TEI e visualizzazione con il software EVT». In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 410-415, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[2] Italia, Paola. *Editing Duemila: per una filologia dei testi digitali*. Roma: Salerno, 2020.

[3] Italia, Paola. «Filologia d'autore digitale». *Ecdotica* 16 (2019): 202–16.

[4] Montale, Eugenio. *Gli ossi di seppia*. A cura di P. Cataldi e F. d'Amely; con saggio di P. V. Mengaldo e uno scritto con S. Solmi. Milano: Mondadori, 2019.

[5] Montale, Eugenio. *L'opera in versi*. A cura di Rossana Bettarini e Gianfranco Contini. Torino: Einaudi, 1980.

[6] Montale, Eugenio. *Tutte le poesie*. A cura di Giorgio Zampa. Milano: Mondadori, 1996.

[7] Nava, Beatrice. «Siamo tutti bédieriani? Prospettive per le edizioni genetiche digitali». *Umanistica Digitale* 6, fasc. 14 (2022): 19–40.

[8] Nava, Beatrice, e Roberta Priore. «Codificare (al)l'Infinito». In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 421-423, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[9] Tancredi, Giulia, e Cristina Fenu. «XML-TEI: Un modello per la filologia d'autore». In *AIUCD 2022 - Proceedings*, a cura di Fabio Ciracì, Giulia Miglietta, e Carola Gatto, 218–22. Lecce: Università del Salento, 2022.

MetaScript: a framework proposal for screenplay encoding

Erica Andreose[1], Giorgia Crosilla[2], Leonardo Zilli[3]
[1] Università di Bologna, Italia - erica.andreose@studio.unibo.it
[2] Università di Bologna, Italia - giorgia.crosilla@studio.unibo.it
[3] Università di Bologna, Italia - leonardo.zilli@studio.unibo.it

ABSTRACT
This paper describes our attempt at developing a framework for the encoding and analysis of cinematic screenplays and the texts that they are adapted from. We propose a set of compliant and reusable methodologies built on top of the existing TEI P5 guidelines, with the aim of providing a systematic approach to portray both the screenplays and the literary texts that inspired them, as well as the intertextual relationships that lie between them and the final audiovisual rendition.

KEYWORDS
TEI; text-encoding; screenplay.

1. INTRODUCTION
We propose in this paper a framework for the encoding of original literary works and their transposition into the audiovisual realm in the form of screenplays. It aims to provide a set of guidelines for the markup and metadata enrichment of texts that span different media, such as books alongside their screenplay adaptation for the big screen. The goal of this framework is to facilitate a more profound analysis of the mediums and their relationship through the alignment of the texts with the final visual rendering of the film, allowing for the comparison of the texts both in their textual and visual forms, supplemented by the inclusion of Linked Open Data metadata and principles.
Film screenplays are a class of literary texts whose analysis and methodologies, especially in the digital realm, have for the most part resembled the ones strictly related to literary studies, not fully considering the visual and cinematic aspects inherent in their transformation from text to screen. Their peculiarity, compared to other types of literary works, is the strictly structured format in which they are written. This format helps the writers of the script to encode a number of various storytelling elements into the text, such as the story, the dialogues and characters' actions, but also more technical information such as camera movements and transitions. We argue that this data, often overlooked in the analysis of literary transpositions, is essential for a comprehensive understanding of how a written narrative is transformed into a visual and auditory experience on screen. This framework aims to bridge the gap between the written word and its filmic interpretation, enabling richer analysis capabilities and exploration of the creative choices made during the process of adaptation.

The examples proposed in this paper are extracted from the project which has been developed as a working prototype for the framework, the subjects being Arthur Schnitzler's novel "Dream Story" [5] and the two screenplays from the film "Eyes Wide Shut" [4] (a draft from 1996 and the official transcription of the film released in 1999), a transposition of Schnitzler's novel directed by Stanley Kubrick and adapted by Kubrick himself along with Frederic Raphael. These source materials will be used to demonstrate the implementation of the presented proposal. All materials are used in compliance with fair use licensing regulations. Textual material usage adheres to research purposes, with only limited portions displayed in the web demonstration of results. For the film, only stills are utilized to capture key frames summarizing analyzed scenes. No video/audio files are reproduced.

2. RATIONALE AND METHODOLOGY
In our methodological approach to the processing of the texts we have made a deliberate choice to adhere as closely as possible to the Guidelines outlined in the official TEI documentation. These guidelines, and more specifically the module for the encoding of "Performance Texts"[1], are already equipped for the encoding of scripts (mainly traditional theatrical texts, with the "Other Types of Performance Text" section describing elements more suited for film screenplays). Therefore, our efforts have been focused on adapting these elements to our needs, introducing new approaches to the encoding of specific elements only in cases in which the characteristics of said elements proved challenging to categorize within the existing TEI framework. In response to these unique requirements, we capitalized on the flexibility of the TEI model,

[1] TEI Performance Texts Guidelines: https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html

encoding the texts with the aim of retaining as much as possible all the crucial information portrayed, while simultaneously enriching them with interconnections to their audiovisual counterparts.
The encoding process has been undertaken for the most part with the aid of scripts that made the task for the most part automated, streamlining the workflow and possibly enabling a broader-scale applicability to possibly other cinematographic screenplays. These scripts, too, have been conceived to be adaptable and modifiable, with detailed documentation to guide users and the option to adjust specific parameters.
This approach not only underscores our commitment to align with established TEI standards but also highlights our dedication to flexibility and extensibility, with a focus on enabling a more efficient and adaptable framework for the treatment of texts connected to some multimedia objects.
As mentioned before, the approach we decided to take in the encoding of the texts has been strongly based on the information provided by the TEI P5 guidelines and more specifically the Performance Texts module, which offers many elements for the encoding of printed dramatic texts and scripts. Elements such as *listPlace* and *listPerson* have been used to list all the geographical locations and the characters that appear in the texts, respectively, assigning a unique identifier to each item. Within the front section, the *castlist* element is utilized to house comprehensive information pertaining to the cast members. Each cast member's particulars are encapsulated within a *castItem* tag. Specifically, the character's appellation is designated by the role element, and a succinct character description is provided through *roleDesc*. To specify the actor portraying the character in the film adaptation, the actor tag is used, with each instance bearing a reference to a unique identifier sourced from IMDb. The information enclosed within the *castlist* element is assembled via an automated Python script, significantly expediting the data collection process. This script harnesses the IMDb API to retrieve the names of all cast members involved in the film adaptation, alongside their corresponding roles and their identifier.
Each scene is denoted by a *div* element, connoting its type and scene number while also being furnished with a distinctive identifier. In cases where a scene correlates with one or more scenes in the transcription, an additional div is incorporated, furnishing information concerning the correspondences with the transcription scenes.
Even for the dialogues contained in the screenplays, the "Performance Texts" module offers appropriate elements such as *stage* and *sp* which, along with their attributes, allow the encoding of scene headings, slug lines, transitions and many other kinds of elements typically found in the formatting of a screenplay.
Going into more detail, the stage directions are meticulously annotated utilizing the stage element and incorporate diverse attributes, including:

- `@environment`, which specifies the contextual setting as being either internal or external.
- `@primary location`, signifying the overarching scene location.
- `@secondary location`, affording a deeper level of specificity, often specifying rooms or buildings.
- `@time`, providing temporal contextualization for the scene.

Furthermore, additional stage directions are included through the stage element, using distinct attributes:

- `@setting`, proffering additional contextual details regarding the scene's setting.
- `@delivery`, employed when the director furnishes guidance on the actor's performance within the scene.

What we couldn't find in the "Performance Texts" module was a way to encode a "bridge" between the screenplay and the audiovisual rendition of it. For this, we resorted to using a timeline element for each speech segment of the screenplay, in which two timestamps are specified to denote the beginning and ending of the line as they are uttered by the actor in the film. These timestamps have been extracted, using automated scripts, from the .srt subtitle file commonly associated with the media file of the film.
For each line of dialogue, the following tagging schema is used:

- The line is enveloped within a `sp` element, which meticulously delineates both the speaker, using the attribute `@who` and the listener, with the attribute `@toWhom`.
- In the transcription, the timeline tag is harnessed to demarcate the start and end running times of the line.
- The `speaker` tag is employed to identify the speaker.
- The actual dialogue line resides within the `p` element.

One of the project's main objectives is to establish connections between the various texts referred to, to provide a comparison and highlight changes in the portrayal of characters and narrative sections. To achieve this, not only the tags for a strictly necessary text encoding are used, but also the addition of div elements is needed to further divide the reference texts for comparison. Regarding character alignment, each character in the texts is included in the *listPerson* tag within the TEI-header and has assigned an identifier. Subsequently, a Macro-XML is created containing these lists of characters and proposing, within the *linkGrp* and *link* tags, the identifiers that refer to the same character in the different analyzed documents. See an example of this in Listing 1.

The alignment of the scenes was done by considering the final script as the reference text, with each scene identified by a unique identifier. The *div* elements were then inserted into the book and the draft screenplay to facilitate the matching of narratological sections, each of which is identified by a unique identifier, along with the corresponding scene number from the transcription. Within the dedicated Macro-XML, *link* is once again used to align the unique identifiers of scenes that correspond across the three texts.

```
<linkGrp>
       <link target="#Fridolin #BILL #BILL"/>
       <link target="#Albertina #ALICE #ALICE"/>
       <link target="#Daughter #HELENA #HELENA"/>
       <link target="#Nachtigall #NICK #NICK"/>
       <link target="#Governess #BABY-SITTER #ROZ"/>
       <link target="#Marianne #MARION #MARION"/>
       <link target="#Gibiser #GIBSON #MILICH"/>
       <link target="#Mizzi #DOMINO #DOMINO"/>
       <link target="#Roediger #CARL #CARL"/>
       <link target="#Pierrette #YOUNG GIRL #DAUGHTER"/>
       <link target="#Gentlemen #JAPANESEMAN1 #JAPANESEMAN2 #KIMONO1 #KIMONO2"/>
</linkGrp>
```

Listing 1. Example extracted from the alignment of the characters

The data modeling process was undertaken to provide a more formalized structure for our project, specifically to gain a better understanding of how each individual element could be interconnected within the broader project. To achieve this, we drew inspiration from the "National Edition of Aldo Moro's Works"[2] and utilized the FaBiO[3] (FRBRaligned Bibliographic Ontology) as the primary ontology to delineate the various levels of Work, Expression, and Manifestation [1: 7] (see Fig. 1):

- *fabio:Work* refers to the general concepts of the book "Dream Story" and of the movie transposition "Eyes Wide Shut".
- *fabio:Expression*, the intellectual content, in our case is the translated version of the book, the draft screenplay and the transcription of the movie.
- *fabio:Manifestation* defines the materialization of the entity, in our case the book version that has been published by Green Integer, the XML-TEI encodings, the transcription of the movie published in 1999 and the draft screenplay.

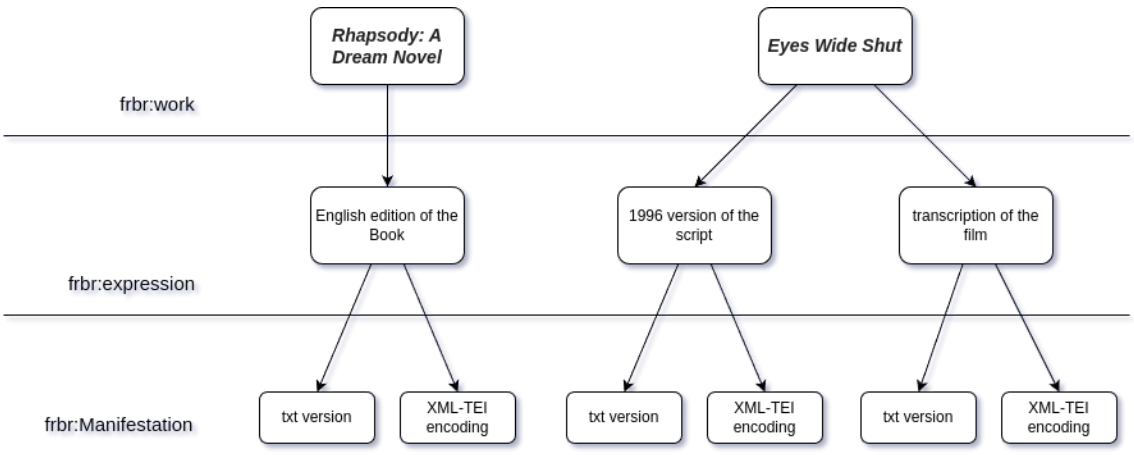


Figure 1. Synthetic data model of the project

The approach followed within the data model was subsequently used to incorporate Linked Open Data into XML-TEI. This was done to propose a shift from the typical document-centric view of XML-TEI, combining it with a more data-centric perspective, which is characteristic of the Semantic Web [2: 50]. To achieve this, RDF triples, expressed through

[2] National Edition of Aldo Moro's works. https://aldomorodigitale.unibo.it/about/docs/models#rdf-section .
[3] Fabio Ontology. https://sparontologies.github.io/fabio/current/fabio.html

subject, predicate and object, were inserted into the *xenoData* tag inside the *TEIheader*, allowing for the inclusion of non-TEI data within the encoding. Furthermore, the *rdf:RDF* tag is used to specify the opening and closing lines of RDF triples. In the opening tag, all ontologies and metadata standards used in the triples are included as attributes. The *rdf:Description rdf:about=""* tag refers to the document, considered here as a *fabio:Item*, as the subject of the triple, followed by additional tags expressing the predicate and object. Moreover, other *xenoData* tags have been added to describe the different FRBR levels, each of those identified with a URI that points to the RDF/OWL document, in which the conceptual structure of the data model has been formally represented. While the predicate is always associated with an authoritative ontology, the object can be a literal or a URI resource derived from a thesaurus or authoritative coding system. Moreover, triples have also been inserted to trace the subject back to its entity expressed in the project's data model and to express relationships between entities. Triples using the predicates *rdau:p60832* ("is inspired by") and *rdau:P60833* ("is inspiration for") were also inserted to describe the internal relationships between the XML files of the original novel and film scripts.

```
<xenoData>
    <rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:owl="http://www.w3.org/2002/07/owl#">
        <rdf:Description
            rdf:about="">
            <rdf:type
                rdf:resource="http://purl.org/spar/fabio/Item" />
            </rdf:Description>
    </rdf:RDF>
</xenoData>
<xenoData>
    <rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:owl="http://www.w3.org/2002/07/owl#"
        xmlns:rico="https://www.ica.org/standards/RiC/ontology#"
        xmlns:frbr="http://purl.org/vocab/frbr/core#"
        xmlns:bibo="http://purl.org/ontology/bibo/"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:fabio="http://purl.org/spar/fabio/"
        xmlns:rdau="http://rdaregistry.info/Elements/u">
       <rdf:Description
rdf:about="https://purl.org/metascript/data/manifestation/draft1996/txt">
        <owl:sameAs rdf:resource="http://www.archiviokubrick.it/opere/film/ews/script/ews-
script.html" />
            <rdf:type rdf:resource="http://purl.org/spar/fabio/Manifestation" />
            <dc:title>Eyes Wide Shut</dc:title>
            <dc:creator rdf:nodeID="creator" />
            <dc:creator rdf:nodeID="creator" />
            <dcterms:date>1996-04-08</dcterms:date>
       <rico:isDraftOf
       rdf:resource="https://purl.org/metascript/data/expression/screen1999"></rico:isDraf
       tOf>
        </rdf:Description>
        <dc:Agent rdf:nodeID="creator">
            <dc:name>Stanley Kubrick </dc:name>
            <rdfs:seeAlso rdf:resource="https://viaf.org/viaf/14772018/" />
        </dc:Agent>
        <dc:Agent rdf:nodeID="creator">
            <dc:name>Frederic Raphael </dc:name>
            <rdfs:seeAlso rdf:resource="http://viaf.org/viaf/24643548" />
        </dc:Agent>
    </rdf:RDF>
</xenoData>
<xenoData>
    <rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
```

```
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:owl="http://www.w3.org/2002/07/owl#">
        <rdf:Description
            rdf:about="https://purl.org/metascript/data/expression/translation">
            <rdf:type
                rdf:resource="http://purl.org/spar/fabio/Expression" />
            </rdf:Description>
    </rdf:RDF>
</xenoData>
<xenoData>
    <rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:owl="http://www.w3.org/2002/07/owl#"
        xmlns:rdau="http://rdaregistry.info/Elements/u">
         <rdf:Description
            rdf:about="https://purl.org/metascript/data/work/eyeswideshut">
            <rdf:type
                rdf:resource="http://purl.org/spar/fabio/Work" />
                <owl:sameAs
                rdf:resource="https://viaf.org/viaf/316753477/" />
            <rdau:P6083>
rdf:resource="https://purl.org/metascript/data/work/traumnovelle"></rdau:P60832>
            </rdf:Description>
    </rdf:RDF>
</xenoData>
```

Listing 2. Example extracted from the project prototype xenoData section.

3. PRELIMINARY RESULTS

The preliminary results are condensed into the draft of the project, which is currently published on GitHub[4] and accessible to the public. In the development of the prototype website, significant attention has been dedicated to the visualization and simultaneous comparison of corresponding scenes in the three different texts under consideration, as well as their visual representation with frames extracted from the film. This allows for an immediate visualization of existing differences and adaptations made by the book in relation to the director's requirements. This was achieved thanks to the modeling choices taken in the encoding phase of the project and XML's querying capabilities using XPath expressions, fetching the portion of the texts corresponding to the scene selected by the user along with the corresponding screencaps which have been collected from the media file of the film, which are also uploaded and aligned in the resource.

Furthermore, a section was created concerning the visualization of data extracted from the texts, particularly focusing on geographical analysis, characters' networks across various texts, and spatio-temporal aspects related to the screenplay [3]. These analyses were performed using various Python libraries and were developed to be as generic as possible, thus reusable to source materials different from those analyzed. In our case study, it became evident through spatial analysis that significant disparities exist between the book and the screenplay. The original European setting underwent a transformation by the director, evolving into a more contemporary New York backdrop. Nevertheless, upon comparing the character relationships network depicted in the book, with those in the 1996 and 1999 screenplays, it becomes evident how the ensemble of characters orbiting the main protagonist is expanded. Lastly, a segment of the spatiotemporal analysis focused on the director's selection of scene settings in the 1999 screenplay. This analysis shows the distribution percentages between external and internal locations, day and night settings, enabling us to visualize the corresponding frames from the movie.

These initial analyses have been valuable in providing a deeper and immediate understanding of the three texts under examination, and particularly in detailing the setting changes between the film and the book, how character names and their interactions change, and which setting is more predominant in the film.

[4] The project was developed for the examination of the course "Digital Text in the Humanities: theories, methodologies and applications" taught by professor Tiziana Mancinelli in the "Digital Humanities and Digital Knowledge" master's degree at University of Bologna. The demo of the project is available at https://giorgiacrosilla.github.io/metascript/index.html. Source files and scripts are available on request.

4. ONGOING AND FUTURE DEVELOPMENTS

Our research allowed us to highlight how, even though there are no TEI guidelines specifically aimed at encoding of screenplays in relation to the visual transposition, the flexibility of the TEI framework allows for the already existing TEI tags to be adopted in order to address these gaps. Since, as of today, this project has only been created and tested on a single set of source materials, future developments would concern the use of different screenplays and texts from which the script is adapted from to verify the effectiveness of the encoding model and determine whether it can be applied to other texts, and if adjustments are needed. Another future objective concerns the expansion of Linked Open Data used in the project, enlarging the network of connections that tie entities within and outside the project. To improve the dialogue and coordination between the script and the textual transposition, it would be necessary to refine the process of extracting and inserting the timestamp references into the encoding. Currently, the timestamps have been automatically extracted from the film's subtitle file, presenting some limitations such as only allowing for the alignment with segments containing dialogues. In instances where important scenes are silent, manual intervention becomes imperative. This approach is inherently restrictive and labor-intensive. In the near future, we intend to embark on the development of an enhanced solution for the extraction of the timestamps designed to comprehensively annotate any scene within the film. These enhancements are directed towards improving the textual encoding framework, with the goal of expanding the scope of data acquisition and analysis achievable during the information extraction phase.

This endeavor envisions the potential for future developments to introduce additional data elements, fostering more comprehensive analyses of character relationships, emotional dynamics, and visual devices (such as the color palette of scenes) employed in the film. Such an enriched framework would enable a deeper analysis of the differences and congruities between the different media formats, facilitating an interesting exploration of the creative choices made in the adaptation process.

REFERENCES

[1] Barzaghi, Sebastian. 'Data Modelling in the National Edition of Aldo Moro's Works (2.0.1)'. *Zenodo*, 2021. https://doi.org/10.5281/zenodo.5524746.

[2] Daquino, Marilena, Francesca Giovannetti, and Francesca Tomasi. 'Linked Data per le edizioni scientifiche digitali. il di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini'. *Umanistica Digitale* 3, no. 7 (2019): 49–75. https://doi.org/10.6092/issn.2532-8816/9091.

[3] Hoyt, Eric, Kevin Ponto, and Roy Carrie. 'Visualizing and Analyzing the Hollywood Screenplay with ScripThreads'. *The Alliance of Digital Humanities Organizations and The Association for Computers and the Humanities* 8, no. 4 (2014).

[4] Kubrick, Stanley, and Frederic Raphael. *Eyes Wide Shut: A Screenplay Dream Story*. New York: Grand Central Pub, 1999.

[5] Schnitzler, Arthur, and Otto Paul Schinnerer. *Dream Story*. Green Integer, 2003.

[6] Tillett, Barbara. 'What Is FRBR? A Conceptual Model for the Bibliographic Universe'. *Technicalities* 25, no. 5 (2003).

OpenData: OpenGadda

Eleonora Pasquale[1], Martina Pensalfini[2]
[1] University of Bologna, Italy - eleonora.pasquale@studio.unibo.it
[2] University of Bologna, Italy - martina.pensalfini@studio.unibo.it

ABSTRACT

OpenGadda is an ambitious project that aims to handle the oftentimes discussed problematic of copyright, in the textual scholarship field, through the employment of all the – freely accessible – data and metadata related to a specific author.
Furthermore, this project will involve the creation of a future paradigm and model that will further showcase the shift from a media-oriented perspective to a data-oriented one and explore the concepts of work and document in the digital era and what they entail in this specific case, the one of Carlo Emilio Gadda, as none of the Gadda's works are currently available freely to the public due to copyright laws.
To obviate this problem, it was chosen to proceed by creating a collection of digital reproductions of different publicly available documents of this author, enhanced through the tools of data visualization and storytelling; specifically, the project focused on the author's archives and library.

KEYWORDS

Textual scholarship; genetic criticism; copyright; digital archives; authorial libraries.

1. INTRODUCTION

Oftentimes the process of digitization of historical artifacts is stalled by copyright which deny access to both the material artifact and the digitalized product.
Unless the artifact is public domain – a work with no intellectual property rights – it still belongs under the scope of exploitation of work with all the inherent features and limits. This entails that oftentimes there are limits to what can be published and used online and not [7].
Two emblematic cases of these issues are the work of Elsa Pereira [8] and the case of *James Joyce's Correspondence*; in the first case, the researcher found out at the last minute that the permission for her research around the figure of the author she was studying had been withheld by the heirs, effectively stopping the project before it could even start. This is a very common outcome in Portugal, due to the importance of moral rights and their permanence[1].
In the second case, instead, access to the resource *James Joyce's Correspondence* was restricted for some countries, as, while Belgium – the country where the project was developed – had by now exhausted James Joyce's copyright protection, different rules and regulations entailed far longer period for different countries, such as Spain and UK.
As shown by these two situations, the absence of clear protocols or an appropriate description of digitization makes it so that not only there is the need to ask for permission through a long chain of communication to create a digital reproduction, but also the digital reproduction itself is oftentimes unable to be accessed due to technical protection measures [6].
Oftentimes various solutions have been offered to deal with this problem; the most infamous and employed by one of the most important digital editions of a 20th-century author – the *Samuel Beckett Digital Manuscript Project* – is to settle a deal with the author's Estate [10]. This allows them to reproduce the documents but under a paywall that is the specific request of the Estate[2].
This does not allow open access to the resource, and neither is applicable in some cases where the whole of the author's work has been split through different Estates and rightsholders, as this would entail an even longer wait for permission and the impossibility to publish a complete collection if also only one of the rightsholders refuses.
Naturally, this brought forward the need for another solution to represent the author and their work and we came up with *OpenGadda*; *OpenGadda*[3] aims to create an open environment, specifically a website, where the user can consult the

[1] *James Joyce's Correspondence.* A project based on the collaboration between the Oxford Centre for Textual Editing and Theory (University of Oxford), the Centre for Manuscript Genetics (University of Antwerp), the University of Tulsa, Western University (Canada), Pomona College (Claremont, California). The editorial team consists of Sabrina Alonso, Josip Batinic, William Brockman, Ronan Crowley, Kevin Dettmar, Michael S. Groden†, Robert Spoo, and Dirk Van Hulle.
https://joyceletters.uantwerpen.be/exist/apps/jjletters/index.html.
[2] *Jame*, is available at the following link: https://www.beckettarchive.org/home.
[3] The website of the project is available at the following link: https://numgadda.github.io/OpenGadda/index.html.

author's work and, through the model of a mental encyclopedia, can collect in one place only all the possible knowledge around the author chosen, employing only freely accessible documents.
The website will collect different materials around the author, all obtained through freely accessible sources, and be divided into separate sections accordingly, on one side offering the user the possibility to browse a catalogue of the items in the collection, while on the other the project will employ the tool of data visualization to visualize the data extracted from the collections and give a further point of view to the user to understand the collection and the author.
Such a new perspective and work plan might create in the future a new paradigm to study an author through new types of documents that differ from the ones classically used in the textual scholarship field, especially in our case we worked on the author's archives and library.

2. A TRANSMEDIAL DIGITAL ARCHIVE

Carlo Emilio Gadda's archives are a prime example for our model, as they offered different challenges and new perspectives upon trying to create a single integrated digital archive; this would give the user the possibility to access all the archives in one single place, with some integrated functionalities to better heighten the experience.
We intended to go a step further from a simple digitization of the archives, but what we aimed was to create a «transmedialized» [9: 22] product, whose focus would not inherently be on the media but instead on the data we worked with. Hence there would be the possibility to visualize the data extracted in different ways either by offering different types of browsing or through the added tool of data visualization [4].
The process through which we created the transmedial digital archive was based on three main operations: data analysis, data visualization and web communication (see Fig. 1).

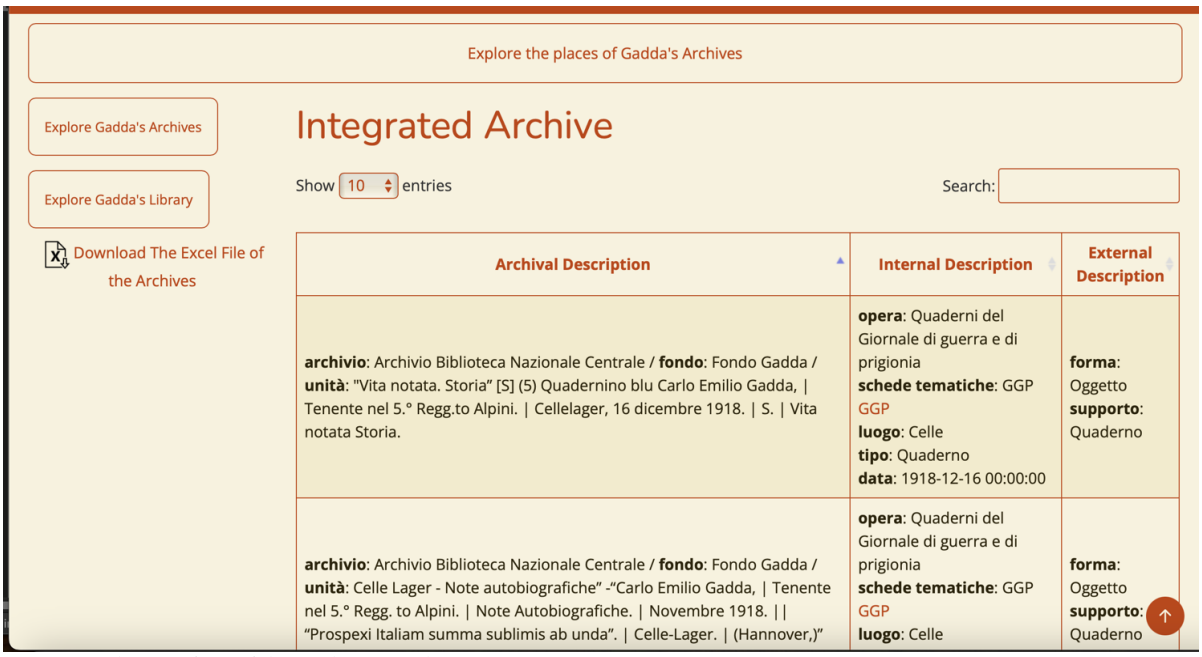


Figure 1. The interface of the Archive Database

During the first phase, we not only extracted the data, but we especially came face-to-face with the various challenges brought forward by literary archives, whose double nature – both literary and archival – has not yet been properly standardized nor cataloged with specific principles. As these archives are both personal and literary, we had to scout through different principles and practices to find the best fitting for the case study that would properly portray the heterogeneous nature of these archives and harmonize the different descriptions present in the different archives [1].
Gadda's archives, especially, proved to be quite complex due to the huge number of records contained in them, alongside to the wide variety; in fact, it was difficult to even come up with the appropriate descriptions and columns for the final database as we aimed to keep up the intricacy of the whole system, which proved as a rather helpful training ground for the modeling of other archives in the future.
When it came to data extraction, it was operated both automatically and manually, as the specific nature of the archives made it quite difficult to just use an automatic retrieval of data. The extraction was based on the source, L'ARCHIVIO DEGLI ARCHIVI, a Word document containing some related descriptions of the archives and their items. After having

created a Word document per archive containing all the information related to it and its items from the source, we transformed it into a .txt file through the PyPandoc Python library.
From here on, we extracted the single units of the archives from the sequence by splitting it at different times for each time the line started again (\n). This respected the actual format that was given in the original document for single units, as they'd be listed one after the other or in numbered lists. Then, there was an attentive employment of Regular Expressions to retrieve the data that composed each unit, once they were split apart, first to attach each unit to their specific series or subseries and then to extract all the possible information around it.
Naturally, as each archive had a different description and format, we had to adapt the extraction algorithms to it, which resulted in different choices and a different final table for each. Also as we said above, there was the need to - once the main part of the extraction was done - refine it through a manual cleaning and filling of the different columns, to ensure a high level of precision and handling a few problems that needed further study.
All the extraction was recorded on Jupyter Notebooks, and all the material obtained by this process was made available on the project's github repository to allow the possibility to reproduce the project in the future.
Secondarily, with the data extracted, we created different visualizations based on the different characteristics of the archives and the author's works and life. Such an operation aimed to create thematic paths that the user could follow and use as a base for their research; the visualization using different tools allowed him to browse the data more immediately and was paired with a brief description that would give a better insight around their context. In this case, the data needed was extracted automatically and the visualizations were created through the employment of the platforms Fluorish and Leaflet.
Three main paths were created (see Fig. 2):

- A geographic one that investigated the role of the main cities in Gadda's life and his archives.
- An analysis of the content of the archives, which was quite interesting due to their heterogeneity; this path, especially, had us wondering about the right approach when handling literary archival documents as we resorted to observing the problem through three different points of view: the ones of the shape, support, and type.
- Finally, we conducted a deep research on thematic cards and their dispersion throughout the archives.

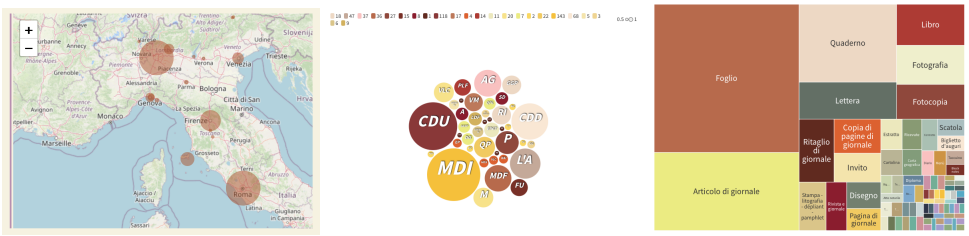


Figure 2. The various visualizations present in the three thematic paths described above

Once we had both the extracted data and visualization, we created a website platform that would contain our dataset in JSON format to be visualized and browsed, through different options (key search, search through the thematic cards or archives, select the published or unpublished works, etc.).
Another interesting feature added to this project was the possibility to connect the archive and library through a table of correspondences between the library genre and the archive thematic cards; by clicking on the library genre, the user is redirected to a library page showing all the results for that specific genre. In this case, we were not able to directly connect the two pages of the archives and library due to problems showing with Internet protocols, but we managed to generate the HTML page dynamically by clicking on the link and showcasing the specific results filtered accordingly to the library genre.
The possibility to browse the dataset in different ways, allowed us to build different layers and methods to analyze and view the content of the archives, offering the users the possibility to interact with it according to their preferences and interests. The data visualization, instead, allowed them to find new thematic research and even encouraged them to create some of their own.

3. A DIGITAL AUTHORIAL LIBRARY

Thanks to the possibility of transgressing the physical limit, libraries nowadays started building themselves on the threshold between library and archive, managing to include a wide variety of materials. This feature is particularly interesting when dealing with the study of authorial libraries, with this term, we intend to identify collections owned by twentieth-century personalities. Even though the concept of digital libraries has not yet been institutionalized and lacks a proper

methodological approach, they are a valuable tool when investigating the personality of an author [3]. The books contained in his library in fact, appear as the "genetic dossier" of his own work, a stratification of both mental and material traces [5: 16].
Nevertheless, the study of a writer's private library and the creation of editions making use of the library itself raise several issues. First, what constitutes the private library: all books a writer owned or those he arguably read? How can we face the challenge of distinguishing between real and virtual library? Dirk Van Hulle, the personality behind the *Samuel Becket Digital Manuscript Project*, proposes a cognitive approach, borrowing Ferrer's idea of genetic criticism, he states that "genetic digital editing may be the key to creating a bridge and a bi-directional exchange between literary studies and cognitive science" [6: 11]. The result is that at the centre of the research, there are the material traces of the writing process. *OpenGadda* proposes a paradigm meant to be easily applicable to other 20th-century writers under copyrights. In this instance, Carlo Emilio Gadda's library [9] allows us to discover a very peculiar study case that collects 2998 volumes, holding in itself a rich tapestry of different genres and authors.
First of all, the information was extracted utilizing the Catalogue of Gadda's Library compiled by Alcini and Giuffrida [2] as the main source. The process saw the usage of the fitz module of PyMuPDF, a high-performance Python library for data extraction of PDF documents. Using an OCR technology we then obtained a txt file and were able to use it as a string. The pieces of information obtained were then converted to a list, that converged into the creation of a database, visualized thanks to the pandas library. The website repository contains the related Jupyter Notebooks utilized throughout the extraction process and is meant to be easily replicable and adaptable to other authors.

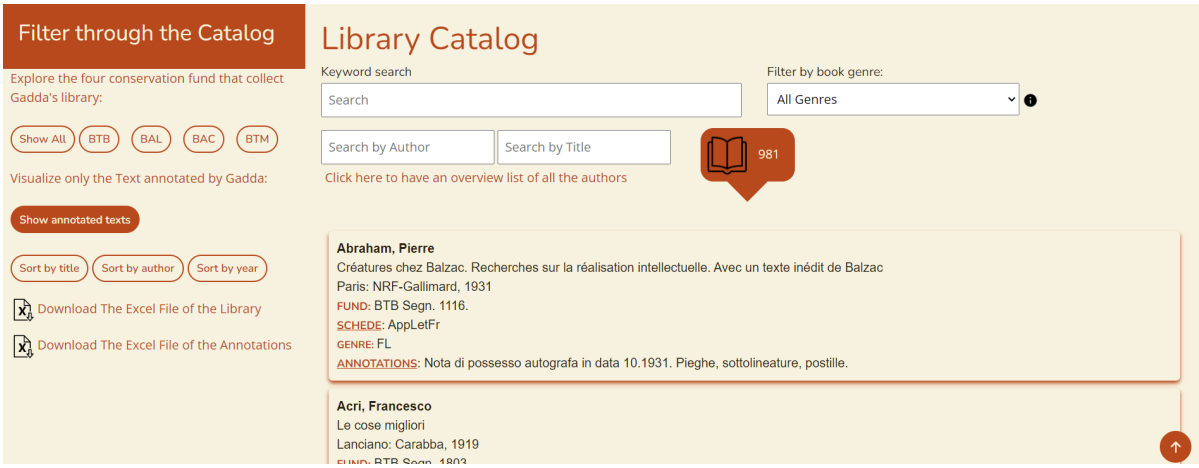


Figure 3. The interface of the Library Dataset

The web environment chosen allows visualization of all the bibliographical entries (see Fig. 3), thanks to the manipulation of the data through a JSON file and uses an Online Public Access Catalogue format; apart from the basic information, it also provides knowledge regarding the conservational fund in which the book is kept and the genre.
When the book happens to contain annotations, these are explicated and are more thoroughly envisionable thanks to a link that opens a dedicated page. The Annotation page allows a straightforward way of visualizing authorial notes, in the case in which there are no photos of the manuscript available. For major clarity, it was decided to distinguish between annotations of ownership, dedications, markings, and proper annotations.
If the author being examined possesses an archive and it is feasible to establish a connection between library items and archival materials, an additional link is generated. This link directs users to the corresponding archival item.
The search engine, besides a generic keyword search entry, offers two separate parameters for authors and book titles, together with a specific filter regarding the book genres.
A sidebar facilitates a more refined research experience by enabling users to filter through conservation funds (or holding institutions). Additionally, it includes another filtering button that allows users to exclusively visualize annotated text. For users seeking a quick overview of all the authors within the library, a dedicated page has been crafted. On this page, each author serves as a direct link to their corresponding bibliographic entries, providing a convenient way to access comprehensive information.
The second section dedicated to the library is a digital storytelling format dedicated to the author's library itself. Data visualization is particularly useful for transforming complex datasets into clear, memorable visuals, facilitating efficient understanding, and communication of insights across diverse target users. To obtain a more immediate comprehension of the visualizations, we chose to create three main thematic areas: authors, books, and annotation (see Fig. 4). Each

visualization is created through technical tools such as Flourish and AmCharts; the ones chosen in the example provided are intentionally tailored to meet the specific requirements of the featured author. However, these visualizations can be customized to suit the preferences and characteristics of any author being examined.

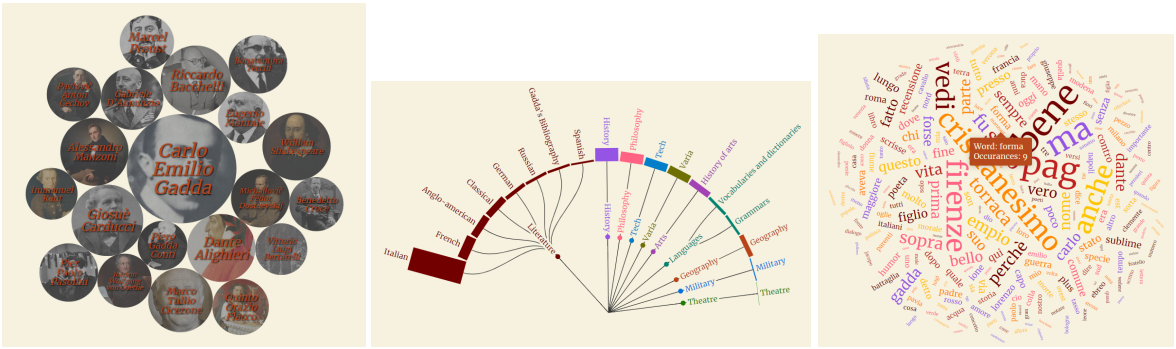


Figure 4. Some of the visualization following the three thematic areas

4. CONCLUSIONS

What we obtained was a digital product that contained all the knowledge stored around the author, not underneath copyright protection, represented both by the browsable dataset that can be consulted in full and the visualization which allowed the user to follow different thematic paths. Obviously, we also instituted a model that could be reproduced in the future with other authors underneath copyright production, as the case study could be applied to different media and different personalities.

To do this we applied the FAIR principles to our project: an appropriate amount of documentation is provided through the website, the datasets are downloadable in an EXCEL format and the source code can be replicated by accessing it from the Github repository[4].

In the future, a wide number of enhancements could be added to improve the digital edition: a downloadable primary bibliography of the author, a comprehensive and up-to-date repository of all the digital resources around the writer and the integration of a LOD system that could expand the network of the project.

Since the project aims to gather all the publicly available documents of the author, if available, the website could be expanded to include other sections dedicated to other materials such as letters, drafts, photographs, and audio recordings.

REFERENCES

[1] Albonico, Simone, and Niccolò Scaffai. *L'autore e il suo archivio*. Milano: Officina Libraria, 2015.

[2] Alcini, Giorgia, and Milena Giuffrida. *Catalogo della biblioteca di Carlo Emilio Gadda*. Roma: Bulzoni, 2022.

[3] Bordalejo, Barbara. 'The Texts We See and the Works We Imagine: The Shift of Focus of Textual Scholarship in the Digital Age'. *Ecdotica* 10, no. 1 (1 December 2013): 64–76.

[4] Décultot, Elisabeth, Paolo D'Iorio, and Daniel Ferrer. *Bibliothèques d'écrivains*. Paris: CNRS Editions, 2001.

[5] Del Vento, Christian. 'Filologia delle biblioteche di scrittori. Come leggeva e postillava Alfieri'. *Autografo* 57 (2017): 39–52.

[6] Driscoll, James, and Elena Pierazzo. 'Introduction. Old Wine in New Bottles?' In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo. Cambridge: Open Book Publishers, 2016.

[7] Klinowski, Mateusz, and Karolina Szafarowicz. 'Digitisation and Sharing of Collections: Museum Practices and Copyright During the COVID-19 Pandemic'. *International Journal for the Semiotics of Law - Revue Internationale de Sèmiotique Juridique* 36 (2023): 1–29. https://doi.org/10.1007/s11196-023-09986-x.

[8] Pereira, Elsa. 'Authors' Rights vs. Textual Scholarship: A Portuguese Overview'. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 14 (2023): 1.

[9] Sahle, Patrick. 'What Is a Scholarly Digital Edition?' In *Digital Scholarly Editing: Theories and Practices*, edited by Elena Pierazzo and Matthew J. Driscoll, 1–22. Cambridge: Open Book Publishers, 2016. https://doi.org/10.11647/OBP.0095.

[10] Van Hulle, Dirk, and Vincent Neyt. 'Developing the Beckett Digital Manuscript Project'. *Wiener Digitale Revue* 1 (2020). https://doi.org/10.25365/wdr-01-03-01.

[4] https://github.com/numgadda/OpenGadda.git

Paul Klee, *Tunisreise* e *Bildnerische Formlehre*: un caso studio di DiScEPT (Digital Scholarly Editions Platform and Aligned Translations)

Hansmichael Hohenegger[1], Tiziana Mancinelli[2], Fabio Ciotti,[3] Federico Boschetti[4], Angelo Mario Del Grosso[5], Eleonora De Longis[6]
[1] Istituto Italiano di Studi Germanici, Italia - hohenegger@studigermanici.it
[2] Istituto Italiano di Studi Germanici, Italia - mancinelli@studigermanici.it
[3] Università degli Studi di Roma "Tor Vergata", Italia - fabio.ciotti@uniroma2.it
[4] CNR Istituto di linguistica computazionale "Antonio Zampolli", Italia - federico.boschetti@ilc.cnr.it
[5] CNR Istituto di linguistica computazionale "Antonio Zampolli", Italia - angelomario.delgrosso@cnr.it
[6] Istituto Italiano di Studi Germanici, Italia - delongis@studigermanici.it

ABSTRACT
Mediante il caso studio dei *Beiträge zur bildnerischen Formlehre* (*Contributi alla teoria figurativa della forma*, 1921-1922; d'ora in poi *Bildnerische Formlehre*) di Paul Klee, il contributo illustra il progetto DiScEPT, piattaforma per la produzione e la pubblicazione di edizioni scientifiche digitali con traduzioni allineate. Uno degli aspetti a cui verrà data particolare attenzione è la possibilità di affiancare le versioni di un testo, o di interi corpora testuali, allineando traduzioni in una o più lingue. Saranno pertanto illustrati i vari componenti della piattaforma all'interno del flusso di lavoro e dei processi di realizzazione di un progetto editoriale e le diverse modalità di modellazione per l'allineamento di traduzioni. Si metteranno in luce le relazioni e le somiglianze tra il nostro pilota della *Bildnerische Formlehre*) e i *Tagebücher 1898-1918* (*Diari*) di Klee, in particolare la parte dedicata al suo viaggio in Tunisia, offrendo così anche un esempio dei modi attraverso cui si possono esplorare i diversi livelli di intertestualità tematica, terminologica e perfino riguardante l'uso delle immagini.

PAROLE CHIAVE
Digital scholarly edition; translation; reuse; TEI; LOD (Linked Open Data); text alignment; stand-off markup.

1. INTRODUZIONE
La piattaforma per edizioni scientifiche digitali DiScEPT nasce con l'obiettivo di raccogliere metodologie, protocolli e buone pratiche consolidate dai molti progetti che sono stati sviluppati nell'ambito della filologia e dell'editoria digitale e che hanno creato una comunità di utenti e sviluppatori intorno a essi. Si vuole dunque riutilizzare e integrare strumenti, servizi e formati già disponibili e ben noti a editori e studiosi in un ecosistema che tenga presente ogni momento della filiera editoriale dalla produzione alla fruizione di un'edizione. L'obiettivo è inoltre quello di avere un ambiente di lavoro per editori più o meno esperti di filologia digitale sulla base dei principi di accessibilità e di riuso [14]. Tre sono i punti in particolare da affrontare:

1) produzione, uso e riuso dei dati e metadati;
2) riuso di componenti software e applicativi e di processi di flusso di lavoro;
3) condivisione di linee guida e buone pratiche.

Questa piattaforma, in primo luogo, permette la produzione di edizioni scientifiche digitali [15], la codifica dei dati e la descrizione dei metadati attraverso formati diversi, come ad esempio XML/TEI, XML-RDF, JSON-LD, focalizzandosi in particolare sull'aspetto dell'allineamento tra edizioni diverse e loro diverse traduzioni. L'obiettivo finale è quello di rendere le pratiche editoriali più rispondenti alle esigenze dei filologi, di creare risorse che abbiano alla base una strategia di sostenibilità e di FAIRificazione dei contenuti. Il fine è di realizzare un ecosistema collegato con altre risorse sul Web e che tenga presente la valorizzazione dei dati come pratica permanente e diffusa nell'ambito degli studi umanistici e del dominio degli archivi e delle biblioteche. Questo ecosistema della testualità digitale si pone dunque come scopo quello di ampliare, attraverso la produzione di dati, e la curatela di questi in ambito editoriale, la conoscenza e l'organizzazione degli artefatti in tutte le loro forme e di ampliare il coinvolgimento del pubblico, nonché di migliorare l'accessibilità, l'inclusività, la fruizione e di rafforzare la ricerca multidisciplinare.
In questo contesto, è bene sottolineare che le traduzioni allineate non sono solo un utile sussidio, ma costituiscono un vero e proprio arricchimento filologico; devono infatti essere considerate un fondamentale strumento per l'analisi

semantica nel suo senso più ampio, non solo nella sua funzione contrastiva e disambiguante. Come nel caso di traduzioni storiche che vengano confrontate con traduzioni moderne, può venire chiamata in causa la storia della lingua, la storia della ricezione e più in generale le varie dinamiche interculturali; anche quando si tratta di traduzioni tra due tipologie di testo diverse, per esempio dalla poesia alla prosa, dal dialogo al riassunto in un discorso indiretto (che richiede naturalmente decisioni tecniche sulla granularità del testo, parola, frase, paragrafo, unità semantica), il confronto può arricchire le analisi stilistiche anche attraverso strumenti computazionali. Proprio per questo motivo l'allineamento [16] prevede che si possano associare sia lingue diverse sia diverse traduzioni nella stessa lingua.
Le due parti, edizione e traduzioni, in DiScEPT non sono pensate come rigidamente separate, ma servono entrambe per studiare la mobilità dei linguaggi, ovvero dei testi che nella propria tradizione/traduzione ne possono rendere testimonianza. In questo senso, e da sempre, la diversità linguistica è una ricchezza per lo studio filologico.
Rinunciamo, però, a sposare una delle tante, magari ottime, teorie traduttologiche o di legarci a una delle tipizzazioni dei testi basate sulle teorie linguistiche di K. Bühler o di R. Jakobson. Piuttosto, senza pretese sistematiche, pensiamo a fornire tipologie di traduzioni basate sui generi letterari, e usiamo i concetti chiave della traduttologia sempre solo operativamente: per esempio, le opposizioni come quella tra traduzione letteraria e traduzione pragmatica [5]; oppure quella tra traduttori *ciblistes* e traduttori *sourciers*, rispettivamente più vicini al testo sorgente o più vicini al testo d'arrivo [13].
Ogni teoria traduttologica viene presa in considerazione se, valorizzabile con un'applicazione digitale, aumenta non solo la qualità ma la stessa coscienza della pratica filologica e/o traduttoria. Così gli strumenti di estrazione della terminologia, le analisi stilistiche comparative o le analisi distribuzionali sono al servizio della migliore risposta alle esigenze dell'editore/traduttore: nel caso quella di rendere più coerente la terminologia (dello stesso o di diversi traduttori) o, al contrario, decidere per una varietà di resa basandosi su un'analisi nella quale il contesto è da ritenersi più determinante.

2. CHE COS'È DiScEPT

DiSCePT è un ecosistema per la creazione di edizioni digitali basato sull'interazione di diversi progetti *open source* e tecnologie aperte che possono gestire differenti tipologie di dati e metadati.
Nel processo di creazione di edizioni scientifiche digitali, si è assistito a un considerevole aumento degli strumenti disponibili, particolarmente marcato negli ultimi vent'anni. Questi sono spesso progettati per facilitare specifiche fasi del flusso editoriale, che includono il supporto alla codifica, l'annotazione di immagini, la visualizzazione e la pubblicazione, come dimostrano esempi quali TEI-Publisher, EVT - Edition Visualization Technology[1] o CETEIcean[2]. Esistono diverse iniziative che forniscono supporto tecnico per la realizzazione di edizioni digitali utilizzando marcature XML/TEI, come *Textual Communities*[3], EVI-LINHD, ecc. Purtroppo, questi progetti in alcuni casi non sono più online oppure sono difficilmente reperibili e non si ha la possibilità di accedere al codice sorgente, limitando così la possibilità di riutilizzare o personalizzare componenti in base a esigenze particolari, come avviene nel progetto DiSCePT per l'allineamento delle traduzioni.
Il gruppo di lavoro ha pianificato la progettazione attraverso una prima fase di raccolta di requisiti caratterizzata da cicli di incontri con filologi tradizionali e con filologi digitali e la stesura di un questionario per delineare gli aspetti e fabbisogni più legati alla critica del testo, alle metodologie, ai metodi e alle funzionalità della piattaforma. I punti su cui il questionario si è focalizzato sono *in primis* le tipologie di attività legate alle esigenze, ai livelli di conoscenza e di esperienza delle persone che idealmente potrebbero usufruire del progetto e operativamente a creare una comunità di utenti. Le sezioni affrontano dunque non solo i bisogni ma anche le questioni più relative alla produzione dei dati, alla loro rappresentazione e elaborazione (vedi il questionario al link: https://tinyurl.com/4ccmxvb2).
L'architettura è formata da diversi componenti e modelli di critica testuale. Il sistema prevede infatti questa macro-struttura:

- un *editor testuale* basato su Domain-Specific Languages (DSL) [1, 2] per facilitare la codifica del testo tramite convenzioni editoriali familiari anche ai filologi tradizionali. Un DSL, facilmente serializzabile in XML con schema proprietario, può essere convertito in XML/TEI tramite fogli di trasformazione XSLT;
- un componente per l'*allineamento di testi bilingui e plurilingui* [3, 5], che produce i dati relativi alla corrispondenza tra l'originale (anche in più versioni) e la sua traduzione (o le sue traduzioni) in un terzo documento, applicando metodologie di stand-off [16, 19] per facilitare la formulazione di molteplici ipotesi di lavoro, a differenti livelli di

[1] https://www.labcd.unipi.it/progetti/evt-edition-visualization-technology/
[2] https://github.com/TEIC/CETEIcean
[3] https://textualcommunities.org/app/

granularità (sezione, paragrafo, enunciato, parola) e per differenti scopi (lettura sinottica, annotazione semantica, analisi linguistica);

- un componente *per la pubblicazione e la visualizzazione*; per questo obiettivo si è scelto di adottare il framework TEI-Publisher[4], a sua volta costruito intorno a eXist-DB (un software open-source la per gestione di database NoSQL in grado di gestire nativamente dati in formato XML). Questo strumento si basa sul concetto di riuso di modelli di documento (ODD - One Document Does it All) implementati tramite TEI processing model[5], e di template di pagina realizzati attraverso una vasta libreria di WebComponent. L'idea di riuso proposta da TEI-Publisher si avvicina molto all'approccio che vuole avere DiScEPT, orientando lo sviluppo del progetto verso componenti riusabili. L'utilizzo di tecnologie come ODD e WebComponent, infatti, renderà possibile condividere porzioni di DiScEPT non solo con comunità con cui siamo già in contatto ma anche con e-editiones - la comunità che gestisce lo sviluppo di TEI-Publisher. La piattaforma DiScEPT si differenzia da TEI-Publisher e progetti simili perché copre la creazione di un'edizione digitale in un processo *end-to-end*, ovvero, offre le funzionalità che vanno dalla creazione della collezione dei documenti TEI contenenti le diverse traduzioni, alla esposizione dei dataset tramite interfaccia web, PDF o print CSS. Questi strumenti permettono la creazione di progetti digitali composti da uno o collezioni di documenti, definiti tramite documenti XML/TEI, risorse RDF [4] e manifesti IIIF. Ogni collezione è composta da documenti che possono essere aggiunti e gestiti tramite una comoda interfaccia web, protetta da autenticazione e possono mappare dei testi sorgenti e le rispettive traduzioni offrendo diverse modalità di modellizzazione, visualizzazione e interazione.
- un componente per l'*annotazione di immagini esposte con il framework IIIF*, che consente al lavoro editoriale di evidenziare, commentare e condividere dettagli specifici delle immagini. In aggiunta, un ulteriore componente promuoverà l'esposizione e l'*arricchimento dei dati utilizzando modelli semantici*. Questo approccio facilita una gestione più dinamica delle informazioni, migliorando l'interoperabilità, la ricerca e l'analisi dei dati scientifici, nonché l'espressività e l'organizzazione dei dati.

3. LA *BILDNERISCHE FORMLEHRE* E I *TAGEBÜCHER* DI PAUL KLEE

Il primo progetto editoriale che sarà ospitato dalla piattaforma DiScEPT è l'edizione digitale della *Bildnerische Formlehre* [7, 8, 9] di Paul Klee; testo autografo del pittore che raccoglie gli appunti didattici del periodo di insegnamento al Bauhaus di Weimar e Dessau tra il 1921 e il 1931. Oltre all'edizione del testo, il progetto intende mostrare le sue potenzialità nella rappresentazione di complesse reti intertestuali, mostrando i diversi tipi di correlazioni significative tra alcune sezioni del testo con i temi che Klee espone nelle parti dei suoi *Tagebücher 1898-1918* [11, 12] (d'ora in poi *Tagebücher*) che riguardano il suo viaggio in Tunisia.

Esiste un'interessante relazione, ancora tutta da studiare e approfondire grazie agli strumenti digitali, tra i quattro quaderni dei *Tagebücher* e il quaderno di appunti per le lezioni tenute al Bauhaus (*Bildnerische Formlehre)*. Il fatto stesso che siano scritti su supporti simili, quaderni di tipo moleskine, e che siano a livello cronologico quasi immediatamente successivi fa intendere la continua necessità di Klee di sviluppare, in un peculiare dialogo tra parole e immagini, la riflessione sull'arte propria e quella insegnabile. Nei *Tagebücher* questo colloquio è più interno e il rimando ai suoi disegni e quadri rappresenta l'intimo processo euristico della pittura architettonica o del colore vissuto dall'autore.

Klee continuerà questa ricerca sulla natura della figurazione con parole e immagini, fedele all'idea che le immagini non siano 'semplici' illustrazioni, ma parte integrante dell'indagine stessa [10][6]. Ai *Tagebücher* segue, pertanto, in modo naturale la *Bildnerische Formlehre.* Dopo il 1923 Klee non scriverà più su quaderni e raccoglierà le proprie riflessioni sulla pittura in fogli sparsi che andranno a comporre il cosiddetto *Pädagogischer Nachlass* (*Lascito pedagogico* [9]). Si tratta di circa 3000 pagine, raggruppate dallo stesso Klee in gruppi tematici. In una ideale pubblicazione di tutti questi testi, sarebbe assai utile mostrare come gli strumenti digitali possano rendere esplicite dinamicità e interne relazioni in questo «piccolo viaggio nel paese di una migliore conoscenza» [10: 28].

Il *Klee Zentrum* di Berna conserva il patrimonio di Paul Klee comprendente i suoi diari, il catalogo ragionato scritto a mano, gli appunti delle sue lezioni al Bauhaus, gran parte della sua corrispondenza con la famiglia, gli amici, i conoscenti, i mercanti d'arte e i musei, oltre a documenti personali e fotografie; inoltre, vi è conservata la biblioteca di Paul e Lily Klee, la collezione di oggetti naturali dell'artista, nonché colori, strumenti di pittura e altri utensili del suo studio. I testi principali oggetto del lavoro sono i manoscritti originali e le prime edizioni storiche [7, 11]. Benché queste siano state importanti per la conoscenza e divulgazione del Klee teorico-pedagogico, sono poco utilizzabili per via di molti interventi

[4] https://teipublisher.com
[5] https://tei-c.org/release/doc/tei-p5-doc/en/html/TD.html#TDPMPM
[6] https://de.wikisource.org/wiki/Sch%C3%B6pferische_Konfession:_Paul_Klee

interpretativi arbitrari. Le due edizioni storico-critiche successive, rispettivamente quella di Glaesemer del 1979 [8] e quella di Kersten del 1988 [12], rappresentano notevoli passi in avanti nella direzione di un testo stabilito in modo rigoroso e utilizzabile in vari livelli. Si è tenuto conto anche dell'ulteriore edizione sul web della *Bildnerische Formlehre* che presenta una resa diplomatica del testo manoscritto. Per avere un riferimento anche all'unica pubblicazione destinata alla scuola del Bauhaus, si è esaminata anche la pubblicazione del riassunto essenzialissimo che lo stesso Klee ha fatto dei suoi appunti della *Bildnerische Formlehre* e pubblicato nella collana del Bauhaus: *Pädagogisches Skizzenbuch* 1925 [9]. Naturalmente saranno studiate le traduzioni storiche italiane, la nuova traduzione basata sulla più recente edizione critica e alcune delle traduzioni disponibili almeno in inglese e francese.
È possibile restituire questa complessa storia editoriale, in modo critico e documentabile, solo con strumenti digitali integrati in modo innovativo.

4. MODELLI DI EDIZIONI E ALLINEAMENTO DI TRADUZIONI

Il caso della *Bildnerische Formlehre* è particolarmente interessante: si tratta, infatti, di un testo che è pensato come didattico, quindi un testo operativo e preciso, ma allo stesso tempo evocativo con forti elementi comunicativi. Si deve inoltre tener conto del peculiare rapporto tra disegno e parola, ovvero di elementi intersemiotici che devono essere conservati nella loro relazione sia nella resa grafica sia nei possibili riusi digitali: nella stessa lingua, ma anche nelle traduzioni.

Le tipologie di edizioni individuate per progetti editoriali con traduzioni possono essere molteplici. In particolare, gli scritti di Paul Klee prevedono redazioni multiple sia dei testi originali sia delle traduzioni. Le nostre scelte di architettura dovranno essere, tuttavia, abbastanza flessibili da consentire di lavorare nel nostro ambiente digitale con altre tipologie: ovvero con casi diversi di tradizione/traduzione, come per la traduzione della poesia [17], oppure per i testi filosofici [6].
L'allineamento è effettuato seguendo le linee guida del modulo 16: "Linking, Segmentation, and Alignment" della TEI version P5 prendendo in considerazione la struttura di *stand-off* attraverso l'elemento `<annotation>`. Questo si ispira e segue appunto la specifica pubblicata nel 2017 dalla W3C *Web Annotation Data Model*. *Annotation* può essere all'interno di un elemento che può contenere e annidare diverse tipologie di annotazioni come `<listAnnotation>`.
La *stand-off annotation* riveste un ruolo cruciale nell'allineamento di traduzioni, offrendo un approccio che conserva la struttura originale dei documenti senza alterazioni, permettendo un confronto diretto tra testi originali e tradotti. Questa metodologia assicura una notevole flessibilità, consentendo collegamenti dettagliati a vari livelli di granularità (sezione, paragrafo, periodo sintattico, parola), dalla macrostruttura ai dettagli più minuti come singole parole o frasi, facilitando così analisi approfondite sulle tecniche di traduzione e sulle scelte stilistiche. L'interoperabilità e la riusabilità delle annotazioni stand-off potenziano la collaborazione tra progetti e sistemi diversi, anche ampliando le possibilità di ricerca interdisciplinare nella creazione delle annotazioni. Questo approccio è fondamentale per l'allineamento di traduzioni perché offre un metodo flessibile e dettagliato per collegare testi a diversi livelli di granularità, migliorando la qualità e la precisione degli studi comparativi e traduttologici.
Gli elementi *stand-off* sono stati scelti anche per la loro capacità di essere utilizzati come contenitori di informazioni esportabili e serializzabili in altri formati quali RDF e JSON-LD. In questo modo, si offre la possibilità di utilizzarne i contenuti in contesti semantici slegati dal XML/TEI. Oltre a questo, visto l'integrazione di IIIF e dell'ultima versione della specifica dell'API Presentation, verrà proposto un modello di *manifest* che possa essere contenitore delle varie annotazioni prodotte sia sulle immagini sia nel testo.

L'obiettivo principale di questo componente è lo sviluppo di modelli e procedure orientate all'edizione scientifica digitale di opere con traduzione e all'usabilità dei documenti prodotti, prestando attenzione sia agli editori, sia agli utenti finali, sia agli sviluppatori di software. Per soddisfare le diverse esigenze, il componente è dotato di vari sottocomponenti configurabili per adattarsi a molteplici scenari di utilizzo e finalità di annotazione. La priorità è garantire che i modelli siano accessibili e fruibili in diversi formati, rendendo così l'esposizione dei dati flessibile e adatta a vari contesti di ricerca e pubblicazione. Questo approccio mira a facilitare il flusso di lavoro, massimizzando l'efficacia nell'analisi critica e nell'annotazione dei testi, oltre a promuovere una collaborazione efficace tra editori, utenti e sviluppatori nella creazione di contenuti digitali arricchiti.
Tutto ciò è possibile grazie a una definizione dei modelli di edizione accurata che utilizzando la potenzialità di Web Annotation Data Model e Stand-off può garantire a edizioni con più redazioni la possibilità di collegare diverse unità testuali.
Le risorse digitali prodotte (dati e strumenti) saranno depositate e rese fruibili tramite i servizi di ILC4CLARIN[7], appartenente al consorzio CLARIN-IT, partner del Progetto H2IOSC (Humanities and cultural Heritage Italian Open

[7] https://ilc4clarin.ilc.cnr.it/

Science Cloud)[8]. ILC4CLARIN è progetto gestito dall'ILC (Istituto di Linguistica Computazionale) parte costituente di DiSCePT. ILC4CLARIN, in quanto centro B di CLARIN, è tenuto a garantire la gestione, la conservazione e la pubblicazione dei dati e dei metadati in modo conforme ai principi FAIR e in vista della long-term preservation (si veda la checklist per i centri B: https://www.clarin.eu/content/checklist-clarin-b-centres).

BIBLIOGRAFIA

[1] Boschetti, Federico, Luca Rigobianco, e Valeria Quochi. «Domain-Specific Languages for Epigraphy: The Case of ItAnt». In *CLARIN Annual Conference Proceedings*, 80–84, 2023.

[2] Boschetti, Federico, Andrea Taddei, Luigi Bambaci, Angelo Mario Del Grosso, Gloria Mugelli, Fahad Khan, e Andrea Bellandi. «Collaborative and Multidisciplinary Annotations of Ancient Texts: the Euporia System». In *The Ancient World Goes Digital Case Studies on Archaeology, Texts, Online Publishing, Digital Archiving, and Preservation*, a cura di Vanessa Bigot Juloux, Alessandro di Ludovico, e Sveta Matskevich. Leiden - Boston: Brill, 2023.

[3] Cushman, Ellen. «Supporting Manuscript Translation in Library and Archival Collections: Toward Decolonial Translation Methods». In *Libraries and Archives in the Digital Age*, (a cura di) Susan L. Mizruchi. London: Palgrave Macmillan, 2020.

[4] Daquino, Marilena, Francesca Giovannetti, e Francesca Tomasi. «Linked Data Per Le Edizioni Scientifiche Digitali. Il Workflow Di Pubblicazione dell'edizione Semantica Del Quaderno Di Appunti Di Paolo Bufalini». *Umanistica Digitale* 3, fasc. 7 (2023). https://doi.org/10.6092/issn.2532-8816/9091.

[5] Froeliger, Nicolas. *Les Noces de l'analogique et du numérique. De la traduction pragmatique*. Paris: Les Belles Lettres, 2013.

[6] Garroni, Emilio. «Kant e il "principio di determinazione" del giudizio estetico». *Paradigmi* VII (1989): 7–19.

[7] Klee, Paul. *Beiträge zur bildnerischen Formlehre*. A cura di J. Spiller. Vol. 1. Basel: Benno Schwabe & Co., 1956.

[8] Klee, Paul. *Beiträge zur bildnerischen Formlehre, Bauhaus Weimar 1921/22*. (a cura di) Jürgen Glaesemer. Basel: Schwabe & Co., 1979.

[9] Klee, Paul. *Pädagogisches Skizzenbuch.* Umschlagentwurf und Typographie von L. Moholy Nagy. Bauhausbuch 02, Dessau. München: Albert Langen Verlag, 1925.

[10] Klee, Paul. «Schöpferische Konfession». In *Tribüne der Kunst und der Zeit. Eine Schriftensammlung*, XIII:28–40. Berlin: Erich Reiss, 1920.

[11] Klee, Paul. *Tagebücher 1898-1918*. (a cura di) Felix Klee. Köln: Verlag M. DuMont Schauberg, 1957.

[12] Klee, Paul. *Tagebücher 1898-1918*. (a cura di) Wolfgang Kersten. Stuttgart/Teufen: Gerd Hatje-Arthur Niggli, 1988.

[13] Ladmiral, Jean-René. *Sourcier ou cibliste*. Paris: Les Belles Lettres, 2014.

[14] Martignano, Chiara. «A Conceptual Model to Encourage the Development and Reuse of Apps for Digital Editions». *Umanistica Digitale* 5, fasc. 10 (2021): 71–88. https://doi.org/10.6092/issn.2532-8816/12620.

[15] Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham, Surrey: Ashgate, 2015.

[16] Pozzo, Riccardo, Timon Gatta, Hansmichael Hohenegger, Jonas Kuhn, Axel Pichler, Marco Turchi, e Josef van Genabith. «Aligning Immanuel Kant's Work and its Translations». In *CLARIN: The Infrastructure for Language Resources*, (a cura di) D. Fišer e A. Witt, 727–46. Berlin, Boston: De Gruyter, 2022. https://doi.org/10.1515/9783110767377-029.

[17] Romanzi, Andrea. «Er jeg Rolf Jacobsen da? Traduzione e appropriazione dell'identità». In *Incroci. Luoghi della creatività e della comunicazione*, (a cura di) M. Gargiulo, 189–204. Roma, 2020.

[18] Spadini, Elena, e Magdalena Turska. «XML-TEI Stand-off Markup: One Step Beyond». *Digital Philology: A Journal of Medieval Cultures* 8 (2019): 225–39. https://doi.org/10.1353/dph.2019.0025.

[19] Viglianti, Raffaele. «Why TEI Stand-off Markup Authoring Needs Simplification». *Journal of the Text Encoding Initiative* 10 (2019). https://doi.org/10.4000/jtei.1838.

[8] https://www.h2iosc.cnr.it/ - finanziato dall'Unione europea NextGenerationEU – PNRR M4C2 - Codice progetto IR0000029 - CUP B63C22000730005

PAVES-e: Per una Hyperedizione dell'opera di Cesare Pavese

Christian D'Agata[1], Angelo Mario Del Grosso[2], Laura Nay[3],
Giuseppe Palazzolo[4], Antonio Sichera[5], Daria Spampinato[6]

[1] Università di Catania, Italia - christian.dagata@unict.it
[2]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - angelomario.delgrosso@cnr.it
[3] Università di Torino, Italia - laura.nay@unito.it
[4] Università di Catania, Italia - giuseppe.palazzolo@unict.it
[5] Università di Catania, Italia - asichera@unict.it
[6]CNR Istituto di Scienze e Tecnologie della Cognizione, Italia - daria.spampinato@cnr.it

ABSTRACT

L'articolo presenta il progetto PAVES-e, finanziato con i fondi PRIN 2022, che intende creare un'edizione-archivio semantica open access, definita Hyperedizione, dell'opera di Cesare Pavese. Essa permetterà a un vasto pubblico di fruire dell'opera pavesiana attraverso un portale nel quale poter consultare le edizioni scientifiche digitali delle poesie e dei romanzi mettendole in relazione con collegamenti semantici all'epistolario e al diario (*Mestiere di vivere*), fruendo inoltre dei manoscritti, dei vocabolari d'autore e del commento multimediale. Il contributo, dopo aver presentato il progetto (distinguendo analiticamente tutte le sue fasi: *DigitalPavese, OntoPavese, PaveseInImmagini, PaveseInTesto, PaveseInParole, AnnotaPavese, BiblioPavese*), si sofferma in particolare su *PaveseInTesto*, discutendo alcune scelte di codifica e di modellizzazione delle edizioni, con alcune proposte sul workflow, sugli editor scelti e sulla interfaccia di visualizzazione.

PAROLE CHIAVE

Edizioni scientifiche digitali semantiche; codifica XML/TEI; archivi digitali d'autore; lessicografia; filologia computazionale.

1. INTRODUZIONE

Il progetto PAVES-e mira a creare una Hyperedizione dell'opera di Cesare Pavese secondo il paradigma dell'Hyperedizione sperimentato in PirandelloNazionale [5], ovvero un'edizione-archivio open-access, flessibile e profondamente integrata, che possa offrire una modalità innovativa di accedere e collegare tra loro i dati sfruttando le potenzialità delle ontologie formali. Pavese è infatti uno dei pochi autori italiani al quale da anni è dedicato un portale: si tratta di Hyperpavese (www.hyperpavese.com), che conserva le carte dell'autore concesse in comodato e suddivise in due sottofondi: Fondo Einaudi (documenti provenienti dall'omonima casa editrice) e Fondo Sini (carte originariamente conservate presso gli eredi). Hyperpavese ospita le carte dell'autore in formato digitale (circa 13.000 documenti) ma attende ancora una digitalizzazione ad alta risoluzione, uno spoglio rigoroso, una sistemazione accurata, un'indicizzazione puntuale, oltre alla possibilità di una libera fruizione. All'archivio cartaceo gli studiosi hanno attinto per la trascrizione e la pubblicazione di alcuni epistolari [12], dell'edizione critica de *Il mestiere di vivere* [10], di alcune traduzioni classiche e moderne (Dughera, Barberi, Cavallini, Pietralunga), mentre all'archivio digitale hanno recentemente guardato Barbarino per *Lavorare Stanca* [2], Grasso per *La luna e i falò* [14], Nay-Tavella per *Prima che il gallo canti* [13] e Sichera-Di Silvestro per l'*Opera poetica* [12], in cui sono state trascritte e annotate migliaia di carte inedite dell'archivio pavesiano. Hyperpavese può fornire materiali utilissimi per un 'Pavese elettronico', un PAVES-e, che dedichi all'opera del grande scrittore un'Hyperedizione, intesa come uno spazio digitale integrato, dove far dialogare testi e prospettive in una modellizzazione [4] che metta al centro il testo come dimensione virtuale di un'esperienza molteplice (filologica, critica, lessicografica), rivolta a utenti e profili diversi (lettori, studenti e studiosi).

2. STATO DELL'ARTE

Allo stato attuale, le biblioteche digitali esistenti in ambito italiano (Progetto Manuzio o Biblioteca Italiana) consentono un accesso immediato ai grandi testi della tradizione letteraria italiana, con modalità di esplorazione efficaci ma semplici, senza alcuna forma di rappresentazione delle interconnessioni semantiche tra opere letterarie e altri materiali documentari (come immagini, lettere, documenti d'archivio e risorse bibliografiche) e senza alcun supporto per l'analisi del testo e la navigazione. I portali letterari più recenti puntano essenzialmente all'offerta di edizioni scientifiche digitali dei testi [8] o

di edizioni delle carte (con trascrizione e varie forme di esplorazione accluse): Manzoni on-line[1], Dante on-line[2], Archivio pascoliano[3] e Digital Vercelli Book[4], dalla cui esperienza è nato il software di visualizzazione Edition Visualization Technology (EVT) [15]. Uno stato dell'arte completo per quanto riguarda le edizioni digitali si trova in Franzini[5] [7] e in Sahle[6]. In questo contesto, l'Edizione Digitale dell'Opera Omnia di Pirandello[7], rappresenta un modello di riferimento in quanto luogo di sintesi di esigenze scientifiche e culturali diverse: dalla pubblicazione di manoscritti e dattiloscritti (con relativa trascrizione) all'edizione digitale dei testi (con annessa sinossi); dalla vocabolarizzazione dei testi all'applicazione didattica. Filologia, lessicografia e critica si integrano in una visione plurivoca e polifunzionale dei dati testuali [16]. Per quanto riguarda le edizioni semantiche, un modello è invece l'edizione delle Lettere di Vespasiano da Bisticci[8] [18] e l'edizione dell'Opera Omnia di Aldo Moro[9], mentre per la codifica di epistolari e postille modello esemplare sono rispettivamente l'edizione Bellini Digital Correspondence [6] e l'edizione delle Postille di Giorgio Bassani [17]. L'organizzazione della conoscenza del progetto è in accordo a prassi ampiamente consolidate [19], mentre l'ontologia computazionale viene sviluppata sfruttando le moderne tecnologie di web semantico [1], nonché modelli concettuali scientificamente comprovati, quali CIDOC-CRM ed ontologie lessicali e reti semantiche come WordNet[10] e BabelNet [9].

3. IL PROGETTO

Il progetto PAVES-e mira allo sviluppo di un'edizione-archivio che, nel rispetto dei principi FAIR (Findable, Accessible, Interoperable, Reusable) [20], costruisca un'interfaccia innovativa (a) e integri, attraverso una solida ontologia (b), le dimensioni filologiche (c, d), lessicografiche (e), critiche (f) e didattiche (g). PAVES-e prevede infatti (vd. Fig. 1):

a. la progettazione e realizzazione di un'interfaccia web intuitiva e user-friendly (*DigitalPavese*);
b. l'organizzazione delle informazioni e l'integrazione tra le risorse, al fine di sviluppare un'ontologia di luoghi, persone, personaggi, organizzazioni, opere comuni a tutto il patrimonio pavesiano (*OntoPavese*);
c. la realizzazione di un archivio digitale, corredato da descrizioni e metadati, delle immagini di manoscritti autografi, dattiloscritti e prime edizioni (*PaveseInImmagini*);
d. l'elaborazione delle edizioni digitali scientifiche dei romanzi e dell'opera poetica, dell'epistolario e del diario, codificati in XML/TEI e lemmatizzati per quanto riguarda le opere letterarie (*PaveseInTesto*);
e. la creazione di un vocabolario e del relativo software di interrogazione, che comprenda i lemmi delle principali opere pavesiane (*PaveseInParole*);
f. la raccolta dell'archivio bibliografico della letteratura secondaria, indicizzata e strutturata e interoperabile (*BiblioPavese*);
g. la messa a punto di un tool di annotazione dei testi in XML/TEI — o la personalizzazione di un *tool* già esistente — che ne permetta un agevole riuso in prospettiva didattica (*AnnotaPavese*).

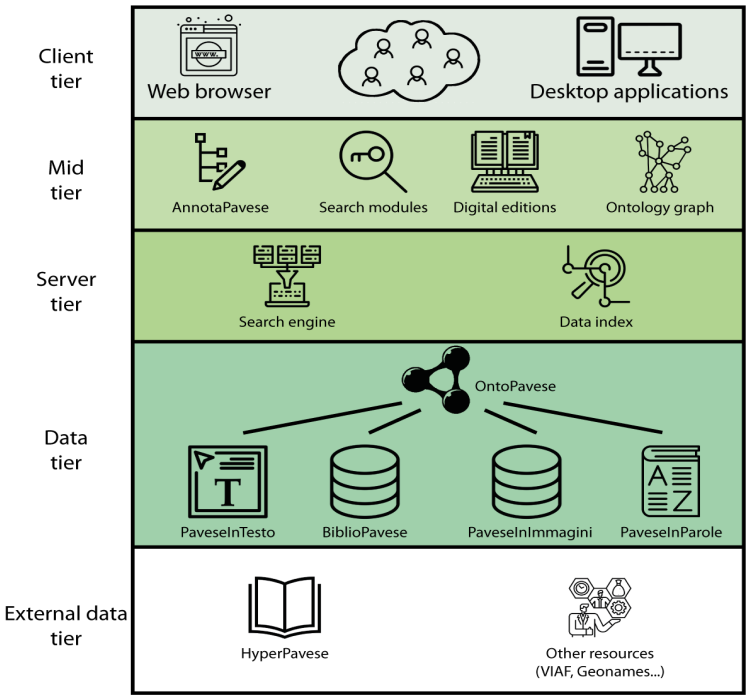


Figura 1. Architettura generale di PAVES-e

[1] https://www.alessandromanzoni.org/
[2] https://www.danteonline.it/
[3] http://pascoli.archivi.beniculturali.it/
[4] http://vbd.humnet.unipi.it/
[5] https://dig-ed-cat.acdh.oeaw.ac.at/
[6] https://v3.digitale-edition.de/
[7] https://www.pirandellonazionale.it/
[8] http://projects.dharc.unibo.it/vespasiano/
[9] https://aldomorodigitale.unibo.it/
[10] https://wordnet.princeton.edu/

4. PAVESEINTESTO: DIGITALIZZAZIONE, ARCHIVIO E CODIFICA DEI TESTI

PaveseInTesto, in particolare, presenterà le Edizioni Scientifiche Digitali dell'opera poetica (*Lavorare stanca, Verrà la morte e avrà i tuoi occhi*), dei romanzi maggiori (*Paesi tuoi, La bella estate, Prima che il gallo canti, La luna e i falò*), dei *Dialoghi con Leucò*, del diario (*Il mestiere di vivere*) e dell'epistolario, a partire dalle edizioni di riferimento già pubblicate o in fase di pubblicazione. Le edizioni saranno allestite attraverso una codifica dell'onomastica, della toponomastica, del discorso diretto e per quanto riguarda le opere letterarie delle correzioni d'autore. Presenteranno un approfondito TeiHeader nel quale saranno esplicitati i metadati fondamentali (ad esempio le informazioni sulla data di stesura delle singole poesie o dei romanzi, le fasi di revisione, la responsabilità di ciascun aspetto della codifica, le informazioni bibliografiche, ecc.). La rappresentazione digitale di tutti questi aspetti sarà effettuata attraverso il sistema di codifica XML/TEI [3], che permetterà di annotare semanticamente i testi integrandoli con l'ontologia di OntoPavese e in modalità Linked Open Data. Per codificare i testi (lettere ed epistolario) il progetto ha previsto l'utilizzo di un tool pensato esplicitamente per la rappresentazione delle relazioni semantiche: LEAF Writer[11] un editor online XML & RDF per il 'Linked Editing Academic Framework'. Questo editor (vd. Fig. 2) permette attraverso una GUI il semplice inserimento di `<persName>`, `<placeName>`, `<orgName>`, `<title>`, `<rs>`, `<quote>`, `<term>`, `<date>`, a partire da un file codificato in TEI (elaborato automaticamente da un semplice script in Python), con la possibilità di aggiungere con grande immediatezza i link a VIAF, Wikidata, DBPedia, GeoNames attraverso un menù a tendina. La scelta è ricaduta su LEAF anche per la possibilità di editare collaborativamente in cloud appoggiandosi su GitHub, in modo da poter avere sempre un *versioning* dei documenti, e contestualmente un parser che validi il documento e permetta di visualizzare sulla destra la codifica 'raw XML'.

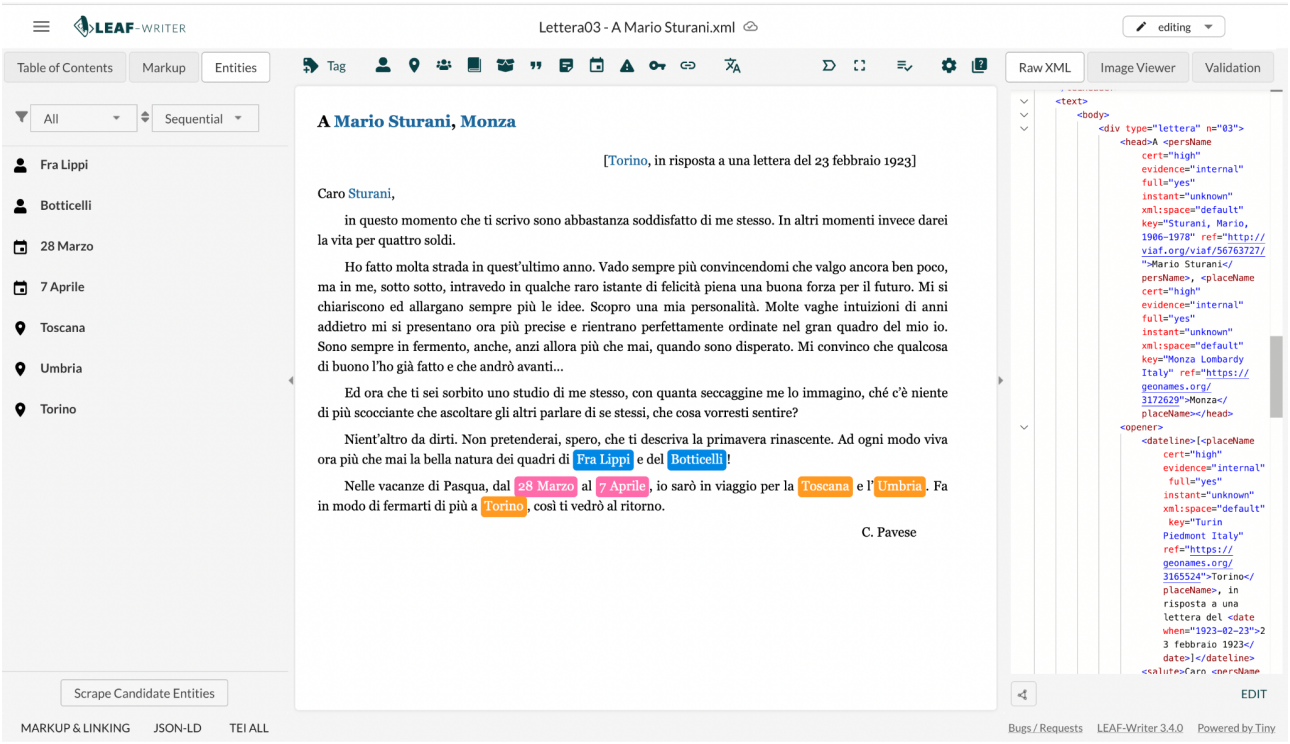


Figura 2. Schermata di visualizzazione del tool di codifica LEAF-writer

Per la codifica dei testimoni e degli apparati critici delle poesie e dei romanzi, ovvero per la rappresentazione dell'aspetto filologico, si è però scelto di utilizzare come editor Oxygen in quanto LEAF è pensato specialmente per la rappresentazione semantica dei Linked Open Data e presenta ancora dei limiti nella rappresentazione degli apparati e, in generale, di tutti quei dati che presentano diversi livelli di gerarchia. Infatti, ad esempio, una poesia come *I mari del Sud* in *Lavorare stanca* presenta tre testimoni (A1, A2 e A3; vd. Fig. 3), di cui uno particolarmente lavorato, com'è anche possibile leggere dall'edizione critica tradizionale (vd. Fig. 4)

[11] https://leaf-writer.leaf-vre.org/

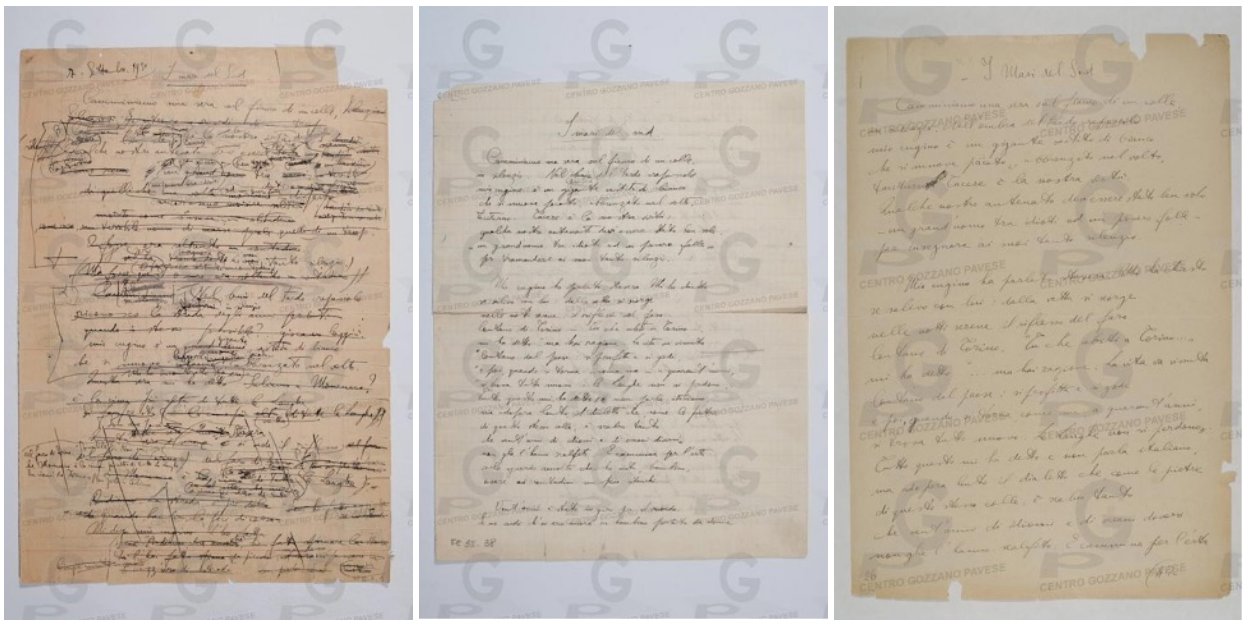


Figura 3. Immagini della digitalizzazione presente su HyperPavese della prima carta delle tre diverse redazioni di Mari del Sud

I MARI DEL SUD

Qualche nostro antenato dev'essere stato ben solo[6]
– un grand'uomo tra idioti o un povero folle –[7]
per insegnare ai suoi tanto silenzio.[8]

Mio cugino ha parlato stasera. Mi ha chiesto[9]
se salivo con lui: dalla vetta si scorge[10]
nelle notti serene il riflesso del faro[11]

6 **A1** Qualche nostro antenato dev'essere [stato ~~un~~ bandito / o uno scemo]
[*stato un*o scemo] *sup.*
[*stato un* bandito] *-sup.*
[*stato un* idiota] *-dx.*
[*stato un* reietto] *inf.*
[morto.] *-sx.sup.*
stato *sup.* ben [triste] *inf.*
[solo] *-inf.*
triste *-dx.*

A2 qualche nostro antenato dev[e] essere stato ben solo
dev'essere

A3 Qualche nostro antenato dev'essere stato ben solo

Figura 4. Edizione critica dei Mari del Sud

Volendo distinguere il momento della rappresentazione del testo da quello della presentazione dei dati attraverso un'interfaccia, ci si è posti il problema su come rendere, innanzitutto nella fase di modellizzazione della codifica, i diversi livelli del testo: quello documentale e quello dell'apparato critico. Tali livelli, infatti, darebbero frutto a due modelli di edizione diversi, l'edizione diplomatica del manoscritto e l'edizione critica con apparato. Per distinguere questi due aspetti si è scelto innanzitutto di rappresentare ciascun testimone di una poesia in `<sourceDoc>`, che è un tagset figlio di `<TEI>` in alternativa a `<facsimile>`, attraverso i tag di aggiunta, cassatura, sostituzione (`<add>`, `<del>`, `<mod>`), inserendo invece nel `<text>` il testo accolto nell'edizione critica. La visualizzazione di ciascun testimone è poi demandata a TeiPublisher[12] (vd. Fig. 5) (opportunamente adattato e implementato per le visualizzazioni degli aspetti semantici e dei Linked Open Data) che permetterà la consultazione contemporanea sia del facsimile digitale delle carte, sia del testo critico, affiancando le informazioni sui personaggi, luoghi, organizzazioni, opere citate.

[12] https://teipublisher.com/

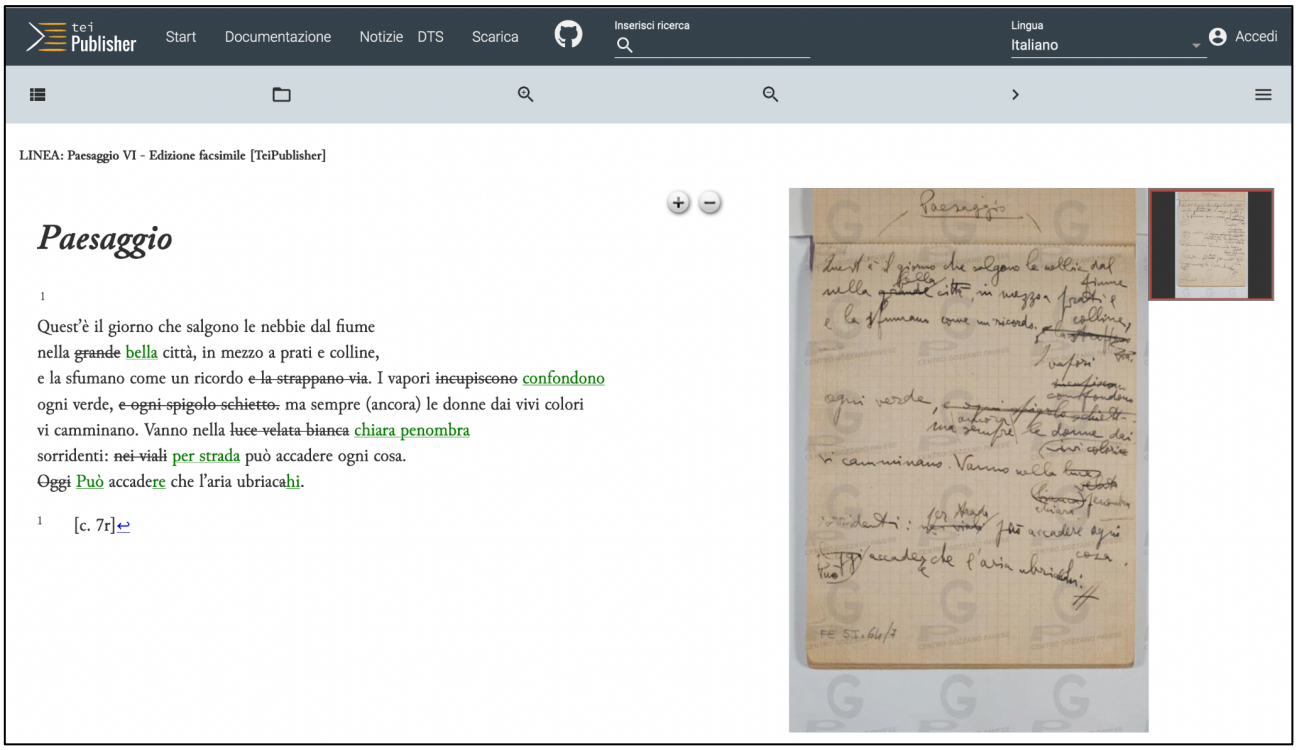


Figura 5. Prototipo di visualizzazione di una carta della poesia Paesaggio [VI] *con TeiPublisher*

5. ULTERIORI SVILUPPI

PaveseInTesto rappresenta certamente *in nuce* le caratteristiche principali di PAVES-e, trattandosi del momento fondamentale di allestimento delle edizioni scientifiche digitali, ma la specificità del progetto e dell'Hyperedizione è quella di prevedere un ampliamento dei tradizionali confini del testo attraverso la lemmatizzazione, il commento multimediale, la rappresentazione semantica dei vari aspetti del progetto, il riuso di tool, lo sviluppo di interfacce ad hoc. In sintesi, a partire dagli elementi già tratteggiati nell'articolo, è attualmente previsto negli sviluppi futuri:

1. Implementazione di un'istanza di LEAF-writer su un server dedicato e sviluppo di funzionalità aggiuntive come la marcatura degli apparati e una migliore gestione dei vari livelli di gerarchia (*AnnotaPavese*).
2. Integrazione e visualizzazione efficace dell'aspetto documentale (`<sourceDoc>`) dei testimoni con l'apparato critico del testo (*PaveseInTesto*).
3. Gestione e collegamento dei dati della lemmatizzazione con la codifica del testo (*PaveseInParole*).
4. Sviluppo di un'interfaccia per la visualizzazione e l'interrogazione dell'ontologia (*OntoPavese*).
5. Sviluppo dell'interfaccia dell'Hyperedizione che permetta di collegare tutti i dati, da quelli filologici, a quelli multimediali e lessicografici, in un'unica modalità di lettura progettata ad hoc per una visualizzazione modulare basata sull'user-centered design (*DigitalPavese*).

6. RINGRAZIAMENTI

Il progetto PAVES-e è finanziato dal Ministero dell'Università e della Ricerca - bando PRIN 2022. Gli autori, inoltre, ringraziano Adriana Damico, Laura Mazzagufo e Alberto Luca Zuliani per il lavoro svolto nell'ambito del progetto.

BIBLIOGRAFIA

[1] Allemang, Dean, e Jim Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Amsterdam: Elsevier, 2011.

[2] Barbarino, Liborio P. *Il primo "Lavorare stanca" di Pavese (1936). Edizione critica*. Avellino: Sinestesie, 2020.

[3] Burnard, Lou. *What Is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources*. Marseille: OpenEdition Press, 2014.

[4] Ciula, Arianna, Øyvind Eide, Cristina Marras, e Patrick Sahle. *Modelling Between Digital and Humanities Thinking in Practice*. Cambridge, UK: Open Book Publisher, 2023. https://doi.org/10.11647/OBP.0369.

[5] D'Agata, Christian, Antonio Di Silvestro, e Antonio Sichera. «Edizione critica, edizione digitale, hyperedizione. "Il fu Mattia Pascal" come paradigma dell'Edizione digitale dell'Opera Omnia di Luigi Pirandello». *Bollettino. Centro di studi filologici e linguistici siciliani* 33 (2022): 263–80.

[6] Del Grosso, Angelo Mario, Erica Capizzi, Salvatore Cristofaro, Maria R. De Luca, Emiliano Giovannetti, Simone Marchi, Graziella Seminara, e Daria Spampinato. «Bellini's Correspondence: a Digital Scholarly Edition for a Multimedia Museum». *Umanistica Digitale* 3, fasc. 7 (2019): 23–47. https://doi.org/10.6092/issn.2532-8816/9162.

[7] Franzini, Greta, Melissa Terras, e Simon Mahony. «A Catalogue of Digital Editions». In *Digital Scholarly Editing: Theories, Models and Methods*, a cura di Elena Pierazzo. Farnham, Surrey: Ashgate, 2015.

[8] Mancinelli, Tiziana, e Elena Pierazzo. *Che cos'è un'edizione scientifica digitale*. Roma: Carocci, 2021.

[9] Missikoff, Michele, Roberto Navigli, e Paola Velardi. «Integrated approach to web ontology learning and engineering». *Computer* 35, fasc. 11 (2002): 60–63.

[10] Pavese, Cesare. *Il mestiere di vivere 1935-1950*. (a cura di) M. Guglielminetti e L. Nay. Torino: Einaudi, 1990.

[11] Pavese, Cesare. *La luna e i falò. Edizione critica*. (a cura di) M. Grasso. Avellino: Sinestesie, 2020.

[12] Pavese, Cesare. *L'Opera poetica. Testi editi, inediti e traduzioni*. (a cura di) A. Sichera e A. Di Silvestro. Milano: Mondadori, 2021.

[13] Pavese, Cesare. *Prima che il gallo canti*. (a cura di) L. Nay e C. Tavella. Milano: BUR, 2021.

[14] Pavese, Cesare, e Bianca Garufi. *Una bellissima coppia discorde. Il carteggio tra Cesare Pavese e Bianca Garufi (1945-1950).* (a cura di) M. Masoero. Firenze: Olschki, 2011.

[15] Rosselli del Turco, Roberto, e Chiara Di Pietro. «La visualizzazione di edizioni digitali con EVT: una soluzione per edizioni diplomatiche e critiche». *Ecdotica* 1, fasc. 2019 (2019): 148–73. https://doi.org/10.7385/99301.

[16] Savoca, Giuseppe. *Lessicografia letteraria e metodo concordanziale*. Firenze: Olschki, 2000.

[17] Siciliano, Angela, e Angelo Mario Del Grosso. «From print to digital: an encoding model for the scholarly edition of Giorgio Bassani's notes». *Umanistica Digitale* 6, fasc. 13 (2022): 21–48. https://doi.org/10.6092/issn.2532-8816/13688.

[18] Tomasi, Francesca, (a cura di). *Vespasiano da Bisticci, Lettere. Knowledge Site 3.0*. Bologna: DH.arc., 2020.

[19] Tomasi, Francesca, Elena Spadini, e Georg Voegler. *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*. Norderstedt: Institut für Dokumentologie und Editorik, 2021.

[20] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Per un'edizione digitale di *Se questo è un uomo*

David Tagliacozzo
Università di Bologna, Italia - tagliacozzo.david@gmail.com

ABSTRACT

L'intervento qui presentato prende in esame il caso di studio offerto da *Se questo è un uomo* di Primo Levi per la messa a punto di un prototipo di edizione digitale del primo capitolo del libro. Allo stato attuale degli studi si possono rintracciare alcune fasi della storia compositiva di *Se questo è un uomo*: le varie redazioni dell'ultimo capitolo conservate in più archivi del Nord Italia, una redazione dattiloscritta di dieci dei sedici capitoli della *princeps*, conservata negli USA, le pubblicazioni su rivista di alcuni brani (cinque sulla rivista vercellese «L'amico del popolo» una sulla fiorentina «Il Ponte»), la *princeps* del 1947 e la seconda edizione, rivista dall'autore, pubblicata da Einaudi nel 1958. Dal confronto tra le varianti di queste fasi emerge un metodo di lavoro peculiare di Levi basato sulla correzione di piccoli particolari e sull'aggiunta di ampi brani, che non vanno tanto a modificare il dettato precedente quanto piuttosto ad approfondire i nessi narrativi e ad aumentare i ritratti di personaggi. Si presenta dunque un prototipo di edizione digitale del primo capitolo di *Se questo è un uomo*, *Il viaggio*, il cui testo è stato marcato nel linguaggio XML/TEI per visualizzarlo tramite il software EVT 2.

PAROLE CHIAVE

Primo Levi; *Se questo è un uomo*; Digital Scholarly Editions; EVT 2.

1. INTRODUZIONE

Il progetto qui presentato riguarda l'edizione digitale di *Il viaggio*, primo capitolo di *Se questo è un uomo* (= SQU) di Primo Levi, libro fondamentale nel panorama letterario europeo – e non solo – sia per il suo valore letterario sia per l'indubbio valore testimoniale e storico che ha assunto, specialmente dopo la seconda edizione per i tipi di Einaudi del 1958. Oltre a illuminare sul metodo di lavoro di Levi, una tale edizione, specialmente se inserita nel panorama digitale, permetterebbe di comprendere il panorama culturale in cui si è mosso il libro: dall'iniziale disinteresse nei confronti della *princeps* (di pubblico più che di critica) – destino condiviso peraltro da gran parte delle testimonianze di sopravvissuti della Shoà (per cui cfr. [13]) – fino al consenso unanime raccolto dall'edizione Einaudi; allo stesso tempo, senza i condizionamenti tipografici legati a un'edizione cartacea, si potrebbe arricchire il testo di caratteristiche ibride, inserendo sezioni di *Passi Paralleli* e *Fonti*, come anche con rimandi a contenuti audiovisivi tramite link.

2. BREVE STORIA DI *SE QUESTO È UN UOMO* DI PRIMO LEVI

Per comprendere al meglio le questioni relative alla creazione del prototipo di edizione digitale del capitolo *Il viaggio* di SQU, creato nel contesto della tesi magistrale in Italianistica, è necessario ripercorrere rapidamente l'iter compositivo del capolavoro leviano.

Tornato a Torino nell'ottobre del '45, Levi iniziò a scrivere il diario degli ultimi dieci giorni nel Lager. Tale resoconto, già allora intitolato *Storia di dieci giorni*, venne terminato nel febbraio del '46, e andrà poi ad assumere la posizione finale di SQU. Il diario venne dunque consegnato all'archivio del Cln, in una copia firmata e datata (= Ist.A)[1]. In seguito Levi ne spedì una seconda – ma si tratta di una copia redatta da un dattilografo esterno – all'archivio della comunità ebraica di Torino (= CET)[2] e altre due, su richiesta di Massimo Adolfo Vitale, all'archivio del Centro di Ricerca Deportati Ebrei di Roma, poi confluito negli archivi del Centro di Documentazione Ebraica Contemporanea (= CDEC1 e CDEC2)[3] di Milano. Nonostante il testo fosse destinato ad archivi di carattere storico, già si può notare una spiccata tendenza letteraria ben diversa dalla semplice testimonianza di fatti atroci, come anche emerge quel carattere di «documento per uno studio pacato di alcuni aspetti dell'animo umano» [9: 3] che sarà poi una delle cifre fondamentali di SQU.

[1] Conservata nel fascicolo a «atrocità nazifasciste», busta C 75 «atrocità nazifasciste», fondi originari, Istituto piemontese per la storia della Resistenza e della società contemporanea 'Giorgio Agosti' a Torino. Subito dopo sono conservate quattro redazioni identiche delle prime quattro pagine del diario (= Ist.B), sicuramente posteriori a Ist.A, di cui accolgono tutte le correzioni.

[2] Conservata nel fascicolo 361, sezione «Relazioni di reduci dai campi di sterminio e denunce (1945)», nell'Archivio della Comunità Ebraica "Benvenuto e Alessandro Terracini" a Torino.

[3] Le due copie si trovano nello stesso fascicolo, il n. 115, busta 3, sottoserie «Vicissitudini di singoli», fondo Massimo Adolfo Vitale, Archivio del CDEC a Milano.

Dal febbraio del ’46 la scrittura di Levi si intensificò: stando ad alcuni dattiloscritti (= AY)[4] inviati alla cugina Anna Yona, residente negli USA, Levi il 14 febbraio 1946 scrisse *Il canto di Ulisse*, il 25 febbraio *Kraus*, a marzo *Esame di chimica*, tra il 5 e l’8 aprile *Ottobre 1944* e tra il 15 e il 20 giugno *Ka-Be*, ma già nel marzo del 1946 Levi aveva esplicitato all’amico Jean Samuel (Pikolo in *Il canto di Ulisse*) l’intenzione di «raccogliere tutto in un libro» [14: 76]. Oltre ai cinque capitoli datati dall’autore (i primi redatti con inchiostro azzurro, l’ultimo con inchiostro nero), questa redazione ne presenta altri cinque, tutti redatti con un inchiostro nero (lo stesso di *Ka-Be*) ma con leggere differenze nell’impaginazione, in particolare nella disposizione dei numeri di pagina[5]. Non è chiaro quale sia stato l’ordine di composizione degli ultimi quattro dattiloscritti (si esclude dal computo *Storia di dieci giorni*, non datato in questa redazione, ma ovviamente scritto prima), ma riflettendo su «l’ordine di urgenza» [9: 3] che informò la scrittura di SQU, è probabile che *Una buona giornata*, caratterizzato da una lieve intenzione positiva, sia stato composto per primo, seguito poi dal trittico iniziatico formato da *Il viaggio*, *Sul fondo* e *Il lavoro*. È infatti possibile intravedere un comune denominatore all’interno dei primi capitoli a essere stati scritti (tra l’altro, quasi tutti concentrati nella seconda metà di SQU) nel loro soffermarsi su effimeri momenti di sospensione del dolore pur nel contesto infernale in cui questi si svolgono. Tutta la redazione AY si caratterizza per una doppia serie di correzioni, una immediata – effettuata direttamente a macchina – e una seconda tardiva – effettuata con una matita rossa con cui Levi corregge gli errori di battitura: sono pochissime le varianti sostanziali che si instaurano in questa fase della composizione di SQU.

Nella ristampa del 1973 di SQU all’interno della collana einaudiana dei «Nuovi Coralli», Levi inserisce gli estremi geografici (Avigliana – Torino) e cronologici (dicembre 1945 – gennaio 1947) della composizione di SQU[6]: a gennaio, quindi, il dattiloscritto venne completato e proposto ad alcune grandi case editrici, tra cui sicuramente Einaudi, ricevendo però un netto rifiuto. Nel frattempo, il 29 marzo 1947, era iniziata la pubblicazione di alcuni capitoli – sottoposti a intensi tagli redazionali – sulla rivista vercellese «L’amico del popolo» (= AP), diretta dall’amico di Levi Silvio Ortona, pubblicazione che però venne interrotta bruscamente il 31 maggio dello stesso anno dopo soli cinque numeri. La successione dei capitoli usciti in rivista è leggermente diversa rispetto all’indice di dell’edizione del 1947: dopo *Il viaggio* e *Sul fondo* (diviso in due parti), viene pubblicato *Le nostre notti* seguito dalla prima parte di *Ka-Be* (col titolo *Un incidente*): *Ka-Be* sembrerebbe avere, infatti, un carattere fluido all’interno del cantiere di SQU (è, tra l’altro, l’unico dei capitoli datati in AY a comparire nella prima sezione del libro): è probabile che inizialmente fosse stato pensato come quarto capitolo e solo più tardi anticipato alla terza posizione per sottolineare la fine del processo di iniziazione raccontato nei primi due capitoli. Il 28 marzo 1947, intanto, il dattiloscritto di SQU (il cui titolo ancora oscillava tra *Sul fondo* e *I sommersi ed i salvati*) venne spedito tramite la sorella di Levi Maria a Franco Antonicelli, direttore della casa editrice De Silva, che accettò con entusiasmo di pubblicare il libro. SQU venne pubblicato nell’ottobre del 1947 (= SQU47), fu il frutto di un intenso lavoro di revisione, concentrata soprattutto nei capitoli scritti per primi. Nello stesso periodo, ad agosto, la rivista fiorentina «Il Ponte» pubblicò il capitolo *Ottobre 1944*.

Molto apprezzato dalla critica ma deludente sotto l’aspetto delle vendite, SQU passò sostanzialmente inosservato: Levi tornò dunque esercitare il mestiere di chimico mentre, segretamente, continuava a pensare a nuovi elementi da inserire nel suo libro primogenito. L’occasione di un ritorno letterario si presentò nel 1955, quando Einaudi accettò di pubblicare una seconda edizione di SQU. L’intensa revisione avvenne sulla copia regalata da Levi alla moglie Lucia (= SQU47c)[7], e consistette in una importante serie di aggiunte orientate a conferire una maggiore patina narrativa all’essenzialità di SQU47 (per cui cfr. [16]) e in una certosina operazione di aumento della precisione linguistica. Quando, nel maggio del 1958, il libro venne stampato (= SQU58), il numero di capitoli era passato da sedici a diciassette, grazie all’inserimento dell’inedito *Iniziazione* a cavallo tra *Sul fondo* e *Ka-Be*; ma anche altri capitoli (in particolare *Il viaggio*, *Sul fondo*, *Le nostre notti* e *L’ultimo*, cfr. [10: 254-255]) vanno incontro a ingenti addizioni testuali, che presentano nuovi personaggi o ne approfondiscono di vecchi: primeggia tra tutte la figura di Alberto Dalla Volta, migliore amico di Levi, cha assurge a un ruolo di comprimario (e doppio del protagonista) nella nuova edizione. Da quel momento la fortuna di SQU è stata ininterrotta: oltre due milioni di copie vendute in Italia e traduzioni, già dal 1959 con l’edizione inglese [6], in quaranta lingue.

[4] Fascicolo conservato nell’archivio dello United States Holocaust Memorial Museum, che ne ha reso disponibile la riproduzione digitale al sito https://collections.ushmm.org/search/catalog/irn538193#?rsc=130713&cv=0&c=0&m=0&s=0&xywh=-1899%2C-1%2C9320%2C6957.

[5] I capitoli in questione sono: *Il viaggio*, *Sul fondo*, *Il lavoro*, *Una buona giornata* e *Storia di dieci giorni*.

[6] Cavaglion [5: vii] e Belpoliti [1: 1456] individuano questo inserimento in SQU58, ma effettivamente non se ne trova traccia fino alla ristampa del 1973.

[7] La lettura di questa redazione, conservata nell’Archivio di Stato di Torino (fascicolo n. 3003, cartella n. 1052) è stata possibile, perciò non è stata collazionata in questo prototipo di edizione.

3. MODELLO DI EDIZIONE

Per il prototipo di edizione del primo capitolo di SQU si è deciso innanzitutto di procedere applicando una marcatura nel linguaggio XML/TEI – ampiamente accettato come standard dalla comunità scientifica, interoperabile e implementabile nel tempo [3: 152] nonostante alcuni limiti[8] – con una marcatura *embedded* (quindi interna al file di trascrizione). La trascrizione è stata effettuata a monte della scelta dell'interfaccia di visualizzazione, per tentare di rendere il documento marcato il più possibile indipendente da strumenti esterni e quindi più facilmente trasportabile.

Come ogni documento in formato XML/TEI, anche questa edizione è strutturata in due macrosezioni, il `<teiHeader>` e il `<text>`: il primo «è il contenitore dei metadati relativi alla pubblicazione, alla descrizione del testo base, ai criteri di edizione e alla dichiarazione dei testimoni utilizzati per l'apparato critico» [7: 218]; il secondo «è il corpo centrale del file Xml, contenente la lezione messa a testo e l'apparato critico» [8: 219].

Dopo la descrizione del progetto e delle sue finalità, nel `teiHeader` sono inseriti: una `<listWit>`, che contiene la descrizione dei testimoni collazionati; una `<listBibl>`, che contiene i riferimenti bibliografici; infine, una `<listPerson>`, per i riferimenti alle persone esistenti nominate in SQU. Per quanto riguarda nello specifico la `<listWit>`, un caso particolare è dato dalla descrizione della redazione AY: pur appartenendo a un unico fascicolo, è molto probabile che i singoli capitoli siano stati scritti in momenti separati, con intenti narrativi via via più chiari allo stesso autore nel corso della costruzione del testo. Ogni capitolo di questa redazione, dunque, è stato descritto come un `<msPart>`[9] all'interno del macrotestimone AY, e descritto secondo le sue peculiarità (inchiostro utilizzato, fasi correttorie, impaginazione, numeri di pagina ecc.).

Il `<text>`, invece, è stato diviso in un `<front>`, che contiene una breve introduzione al progetto e una breve descrizione dei criteri adottati nell'edizione, un `<body>`, che contiene il testo di SQU diviso nei vari capitoli tramite una serie di elementi `<div>`, e un `<back>` che contiene i *Passi Paralleli*.

Per la scelta del testo base (nonostante l'ambiente digitale permetta di rappresentare contemporaneamente più testimoni) si è optato per quello di SQU58, per il suo valore storico (è la versione del testo più nota) e critico (per l'importanza dei dettagli aggiunti a tale edizione, nonostante, come si è detto, questi tendano non tanto a modificare il già esistente quanto a dare nuovi particolari prima accennati o non presenti). Il metodo di codifica dell'apparato è quello della «*Parallel-Segmentation*»[10], il quale prevede che, all'interno del testo stesso, la lezione coinvolta in variante sia marcata all'interno di un elemento `<app>` dentro il quale l'elemento `<lem>` registra la lezione di SQU58 (e dei testimoni che portano la stessa lezione), mentre in uno o più elementi `<rdg>` sono registrate le lezioni precedenti di AY, AP e SQU47. Un tale sistema di marcatura ha, tra l'altro, il vantaggio di essere facilmente convertibile nel *Double End-Point Attachment Method* e di poter annidare le varianti l'una all'interno dell'altra[11]. Per quanto riguarda AY, si è scelto di correggere senza segnalarli tutti gli errori di battitura, sia quelli non corretti da Levi sia quelli emendati in una immediata campagna correttoria (a differenza degli errori presenti nelle copie manoscritte, infatti, gli errori di battitura non portano alcuna notizia sulla lingua dell'autore, cfr. [7]); le varianti interne al dattiloscritto, invece, sono state marcate in un elemento `<app>` in cui la lezione di AY è riportata all'interno di un elemento `<subst>`, dentro al quale l'elemento `<del>` porta la lezione originaria, e un elemento `<add>` contiene la correzione autografa[12]. Nella Figura 1 si trova una schematizzazione del modello di marcatura utilizzato per l'apparato dell'incipit de *Il viaggio*.

Un'alternativa percorribile all'utilizzo dell'elemento `<subst>` per la marcatura delle varianti interne ad AY sarebbe stata quella di registrare, all'interno della `<listWit>` del teiHeader, più `<witness>` per ogni fase correttoria all'interno del fascicolo, secondo il modello sviluppato dal Progetto VaSto, che ha duplicato il testimone su cui si basa l'edizione secondo le due volontà (di autore e di censore) rintracciate [3]. Nonostante il fascino che suscita una tale soluzione, si è deciso di evitare la moltiplicazione dei testimoni, data l'esiguità degli interventi di variantistica rintracciabili all'interno del testimone, a favore di una soluzione certamente più diplomatica ma al contempo più economica. A tutti gli elementi `<lem>` e `<rdg>` è stato poi associato un attributo `@type`, il cui valore indica il tipo di variante (Aggiunta, Taglio, Ortografia, Grafia, Struttura, Stilistica e Linguistica) e un attributo `@varSeq` che indica la successione delle varianti. Nel caso dei

[8] Per cui cfr. [15].

[9] Ogni capitolo della redazione AY è dunque trattato come un singolo testimone all'interno di un fascicolo unito a posteriori: questa soluzione, infatti, da una parte salvaguarda il carattere unitario di AY, dall'altra permette di chiarire le varie fasi della scrittura individuabili al suo interno, marcare ogni capitolo quasi come un testimone a sé, con tutte le proprie particolarità.

[10] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html: 13.2.3.

[11] *Ibidem*.

[12] Per evitare una deriva eccessivamente diplomatica, all'interno degli elementi `<del>` e `<add>` non è inserita solo la porzione del testo cassata e aggiunta, ma l'intera lezione coinvolta in variante.

tagli redazionali rintracciabili in AP è stata esplicitata anche la causa (tramite un attributo `@cause="Redazionale"`) e la responsabilità dell'intervento attribuito a Silvio Ortona, redattore della rivista, con l'attributo `@resp="#S.O."`. Per marcare le interruzioni di pagina è stato utilizzato l'elemento `<pb/>`, a cui è stato dato un identificativo unico (`@xml:id`), un numero (`@n`), un riferimento all'edizione in cui si trova il cambio (`@edRef`) e la posizione della relativa riproduzione fotografica all'interno della cartella di EVT 2; accanto a questo è stato poi inserito un elemento `<lb n="X"/>` per esplicitare il numero di pagina una volta caricato il file su EVT 2.

```
<app>
        <lem wit="#SQU58 #SQU47 #AY" varSeq="2" type="Taglio">
                <app>
                        <lem wit="#SQU58" type="Aggiunta" varSeq="2">
                        <p>Testo dell'incipit di SQU58</p></lem>
                         <rdg wit="#SQU47 #AY" varSeq="1"/>
                </app>
                <p>Resto del testo comune tra SQU58, SQU47 e AY
                <app>
                        <rdg wit="AY" varSeq="1"><subst><del hand="#Penna">lezione
                        cassata</del><add hand="#Penna">nuova lezione in AY</add></subst>
                        </p>
        </lem>
        <rdg wit="#AP1" type="Taglio" cause="Redazionale" resp="#S.O. " varSeq="1"/>
</app>
```

Figura 1. Modello di marcatura del testo e dell'apparato

All'interno del testo sono stati mercati con *tag* appositi anche altri elementi, come i nomi di persona (`<persName>`, con riferimento alla `<persList>` contenuta nel `teiHeader`), le date (`<date>`) e i luoghi (`<placeName>`), così come anche i cambiamenti di lingua (tramite l'elemento `<foreign>` o l'attributo `@xml:lang`).
Un ultimo (doppio) livello di marcatura è stato poi applicato ai passi che rimandano ad altri luoghi dell'opera di Primo Levi (i *Passi Paralleli*) e a quelli che si riferiscono ad altre opere, sia quelle citate da Levi sia quelle in cui l'autore appare come personaggio all'interno di narrazioni del Lager altrui, come nei casi di Jean Samuel o di Luciana Nissim (le *Fonti*). I primi sono stati inseriti all'interno di un elemento `<seg source="X">`, con un collegamento al `<back>` in cui è inserita una lista di tali passi con la citazione del testo. Le *Fonti*, invece, sono marcate all'interno del testo stesso, senza alcun rimando al `<back>`: si è dunque aperto un elemento `<quote>` al cui interno si apre poi un elemento `<bibl>` per il riferimento alla bibliografia contenuta nel `teiHeader`, e un elemento `<quote>` con la citazione del passo: questa possibilità, garantita dal software EVT 2, permette di apprezzare un'edizione non solamente di stampo filologico, ma anche di offrire ulteriori possibilità di lettura, sia da un punto di vista storico-critico sia da un punto di vista contestuale, facendo dialogare SQU con opere coeve, precedenti e successive, inserendolo quindi nel contesto più ampio della produzione leviana e della letteratura concentrazionaria in genere.
Nella prospettiva di un approfondimento delle possibilità di un tale modello di edizione, sarebbe interessante implementare il dossier di materiali messi a disposizione dei lettori: per esempio inserendo una sezione di commento critico e filologico al testo, oppure una sezione in cui poter leggere le recensioni a SQU, specialmente quelle all'edizione De Silva: in questo modo sarebbe possibile apprezzare una delle questioni più controverse della tradizione di SQU, ovvero il dislivello più che notevole tra l'apprezzamento da parte della critica (e di alcuni grandi lettori, tra cui Saba e Calvino) e l'indifferenza del grande pubblico alla testimonianza di Levi.

4. L'INTERFACCIA DI VISUALIZZAZIONE

La scelta di EVT 2 come interfaccia di visualizzazione dell'edizione è dipesa da molteplici fattori: la sua natura «*Prêt-à-Porter*» [11] permette di avere a disposizione gratuitamente uno strumento che sia al contempo semplice da utilizzare (sia da parte del filologo sia da parte dell'utente) ma anche abbastanza flessibile da permettere al curatore di personalizzare l'interfaccia con relativa facilità, per ottenere un'edizione il più possibile vicina ai suoi intenti ecdotici:

> Tramite semplici modifiche al file di configurazione, che consente di attivare o disattivare varie funzioni dell'interfaccia, è infatti possibile accedere ad alti livelli di personalizzazione. La sua natura modulare, che si basa primariamente sulla possibilità di modificare e personalizzare in modo semplice i file al suo interno, lo rende malleabile e adattivo, permettendo di lavorare su più livelli di rappresentazione: sia contenutistica, attraverso il file.xml, che, sebbene in minima parte, di design grafico, attraverso il foglio di stile. [3: 159]

Una volta personalizzato il file di configurazione (config.json) e il foglio di stile (custom-style.css) secondo le necessità di questo prototipo di edizione[13] è possibile visualizzare l'interfaccia EVT 2: questa presenta come prima schermata la modalità «Testo di lettura» (vd. Fig. 2), dove è possibile leggere il testo de *Íl viaggio* nella lezione di SQU58, all'interno del quale sono evidenziate le varianti rispetto ad AY, AP e SQU47: ogni tipologia di variante, secondo il valore dato l'attributo `@type` nella marcatura XML, è caratterizzata da un colore, per rendere subito evidente il metodo di correzione utilizzato da Levi. Accanto al testo è presente un box in cui è possibile leggere i *Passi Paralleli* e le *Fonti*. Cliccando sul tasto *Info* compare una breve descrizione del progetto e dei criteri utilizzati per la trascrizione, mentre cliccando sul tasto alla sua destra si visualizza la legenda dei colori utilizzati per evidenziare i diversi tipi di varianti. Purtroppo, EVT 2 non sembra leggere in qualche occasione l'elemento `<lb/>`, sia all'interno del box delle *Fonti*, sia all'interno dell'apparato, utilizzato nel caso in cui una redazione del testo presenti una diversa struttura dei paragrafi: di conseguenza il testo attualmente non combacia perfettamente con SQU58. Cliccando sulle lezioni evidenziate si apre un box di apparato genetico in cui la lezione a testo (seguita dai testimoni che la riportano e da una parentesi quadra chiusa) è confrontata con le lezioni delle redazioni precedenti.

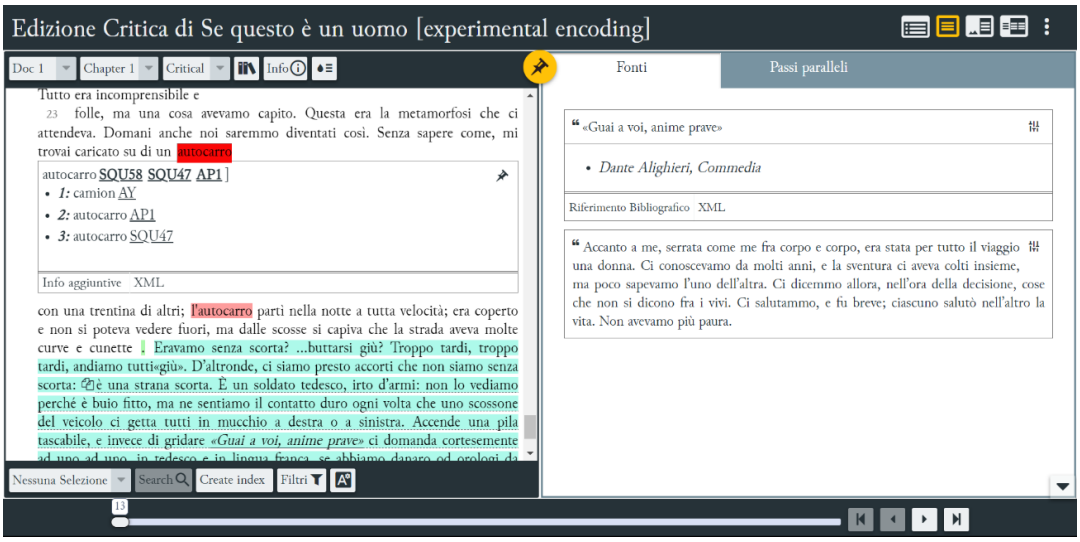


Figura 2. Visualizzazione del Testo di lettura con accanto box delle Fonti e dei Passi Paralleli

Una seconda possibilità di visualizzare l'edizione è la modalità Testo-Immagine (vd. Fig. 3), che permette di leggere SQU58 accanto alla scannerizzazione delle altre edizioni.

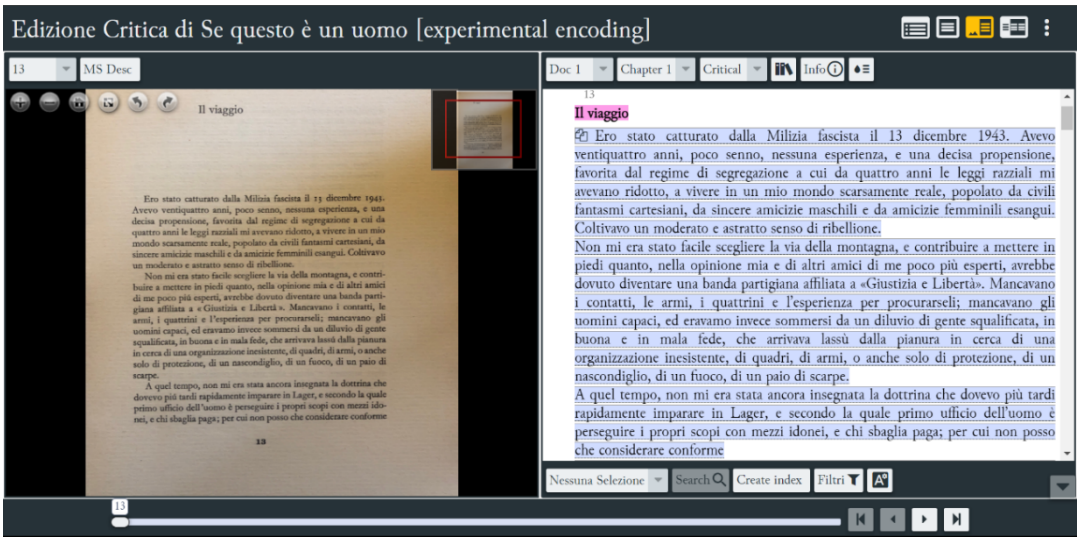


Figura 3. Visualizzazione in modalità Testo - Immagine

[13] Per esempio, per far sì che il testo contenuto nell'elemento `<del>`, all'interno del box di apparato sia reso con un barrato.

Il problema fondamentale di questa modalità di visualizzazione è che non è possibile scegliere di visualizzare immagini del testo di altre redazioni (per esempio AY) e accanto la relativa trascrizione; anche se sembra che la nuova versione del programma miri a risolvere questo problema (cfr. [4]). Ancora, EVT 2, nelle edizioni critiche, non permette lo scorrimento parallelo del testo e dell'immagine: dunque, al cambio di pagina nel box immagini, il testo non si allinea, obbligando l'utente a cercare manualmente la pagina di cui sta visualizzando il facsimile. Un ultimo problema di questa modalità di visualizzazione è nel tasto *Ms Desc*, aprendo il quale si visualizza la descrizione non di SQU58, ma solamente quella di AY.

L'ultima modalità di visualizzazione prevista in questo prototipo di edizione è quella di Collazione (vd. Fig. 4), che permette di confrontare contemporaneamente, in una lettura sinottica, tutte le redazioni che compongono l'edizione digitale, con un apparato che, oltre a elencare le varianti tra i testi, permette anche di visualizzare la marcatura utilizzata tramite il tasto *XML*, e alcune informazioni più precise riguardo la variante, come la sequenza nel flusso degli interventi (*varSeq*), la tipologia di variante in questione (*type*) e i testimoni che la riportano (*wit*), tramite il tasto *Info aggiuntive*. I bottoni nella fascia inferiore del testo, invece, permettono di applicare alcuni filtri al testo (per le *Named Entities*, i luoghi, le date e le tipologie di varianti) e di creare un indice tramite la funzione *Pin*.

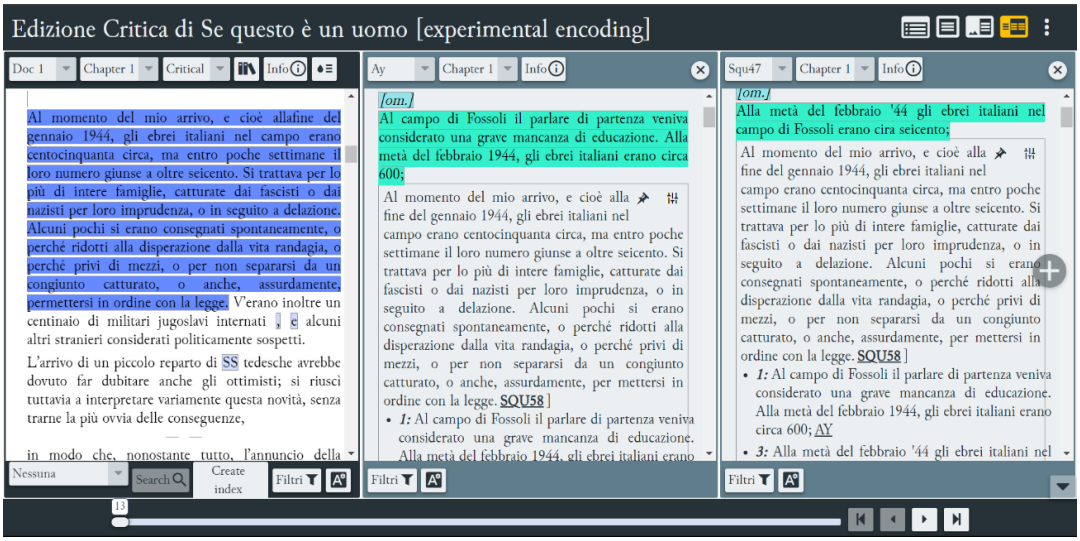


Figura 4. Visualizzazione nella modalità Collazione

Questa modalità presenta relativamente poche problematiche: gli unici limiti riguardano il già citato problema della visualizzazione del tag `<lb/>` all'interno delle voci di apparato e il riferimento alla divisione per pagine delle relative edizioni: EVT 2, infatti, non sembra leggere l'attributo `@edRef` contenuto nell'elemento vuoto `<pb/>`: tale attributo infatti dovrebbe permettere di inserire un collegamento univoco tra l'elemento che lo contiene e i testimoni descritti nella `<listWit>`, ma caricando il documento marcato su EVT 2, ogni cambiamento di pagina è riportato indiscriminatamente su ogni testimone che si sceglie di visualizzare.

5. CONCLUSIONI

Al netto di alcuni problemi riscontrati nella gestione di EVT 2, il risultato raggiunto sembra abbastanza soddisfacente, nonostante manchino ancora alcuni testimoni da collazionare (SQU47c, il quaderno *Per Einaudi* di cui parla Tesio [16: 272], le bozze per le stampe ed eventuali manoscritti preparatori) e non sia ancora stata implementata con rimandi a documenti anche video e audio. Un'edizione digitale di SQU (che sia contemporaneamente critica, commentata e corredata di molteplici filtri interpretativi[14]) potrebbe essere utile nella prospettiva degli studi di filologia digitale: da una parte è un testo relativamente semplice dal punto di vista della sua storia elaborativa, dall'altra offre nella sua semplicità una serie di problemi legati a vari aspetti della filologia, soprattutto se inseriti nel contesto digitale: le questioni "Lachmanniane" della intricata tradizione dell'ultimo capitolo; alcune problematiche legate alla filologia delle strutture (lo statuto fluido di *Ka-Be* o il cambiamento degli equilibri dato dall'inserimento di *Iniziazione*); le questioni sulla stratificazioni di mani che hanno lavorato sul testo (di Silvio Ortona, di Franco Antonicelli, dell'anonimo redattore einaudiano, di Levi stesso). Un'edizione digitale di SQU potrebbe inoltre essere utile e affascinante per una vasta gamma di potenziali lettori: dagli studenti delle

[14] Per cui cfr. [11: 180], per la sua definizione di «Edizione arricchita».

Scuole Superiori che si approcciano per la prima volta a questo «libro fatale» (secondo la felice formula di Saba), agli studenti universitari come anche ai ricercatori specializzati, senza escludere i semplici lettori curiosi di leggere un libro tanto importante per la cultura europea e mondiale. La possibilità di leggere un testo emendato da alcune corruttele della tradizione (come l'anteposizione della poesia *Shemà* rispetto alla *Prefazione*, per cui cfr. [2: 22]) e di poterlo analizzare nelle varie vesti che ha assunto nel tempo (dalle prime bozze fino all'ultima edizione) permettendo non solo di comprendere il modo particolare con cui Levi lavorò sul suo primo libro e sul macrotesto che da quello prende vita, ma anche di illuminare il contesto (letterario e culturale) in cui SQU venne concepito, ignorato e finalmente apprezzato unanimemente.

BIBLIOGRAFIA

[1] Belpoliti, Marco. «Nota al testo di *Se questo è un uomo*». In *Opere complete*, di Primo Levi, 1449–86, (a cura di) Marco Belpoliti. Torino: Einaudi, 2016.

[2] Bersani, Mauro. «La pax del monopolio o il litigio dei filologi?» In *Editori e filologi: per una filologia editoriale*, (a cura di) Giorgio Pinotti e Paola Italia, 21–23. Roma: Bulzoni, 2014.

[3] Brancato, Dario, Milena Corbellini, Paola Italia, Valentina Pasqual, e Roberta Priore. «VaSto: un'edizione digitale interdisciplinare». *magazén* 1 (2021): 139–69.

[4] Cacioli, Erica, Chiara Di Pietro, Sara Maenza, Roberto Rosselli Del Turco, e Simone Zenzaro. «There and back again: what to expect in the next EVT version». In *AIUCD 2022 - Culture digitali. Intersezioni: filosofia, arti, media. Proceedings della 11a conferenza nazionale, Lecce, 2022*, (a cura di) Fabio Ciracì, Giulia Miglietta, e Carola Gatto, 212–17. Lecce: Quaderni di Umanistica Digitale, 2022.

[5] Cavaglion, Alberto. «Presentazione». In *Se questo è un uomo*, di Primo Levi, vii–xiv. (a cura di) Alberto Cavaglion. Torino: Einaudi, 2012.

[6] Farrel, Joseph. «Primo Levi in Great Britain». In Diffusione e conoscenza di Primo Levi nei paesi europei: la manutenzione della memoria: atti del convegno, Torino 9-10-11 ottobre 2003, (a cura di) Giovanni Tesio, 107–37. Torino: Centro Studi Piemontesi, 2005.

[7] Italia, Paola. *Editing Duemila*. Roma: Salerno, 2020.

[8] Italia, Paola. «Il testimone anfibio». In *La tradizione dei testi. Atti del Convegno Cortona, 21-23 settembre 2017*, (a cura di) Claudio Ciociola e Claudio Vela, 253–75. Firenze: Società dei Filologi della Letteratura Italiana, 2017.

[9] Levi, Primo. *Se questo è un uomo*. (a cura di) Alberto Cavaglion. Torino: Einaudi, 2012.

[10] Pepe, Tommaso. «Genesi di *Se questo è un uomo* e autoriscritture leviane del Lager». *Misure Critiche* XV, fasc. 1–2 (2016): 225–58.

[11] Pierazzo, Elena. «Edizione documentaria digitale: rinuncia intellettuale o opportunità scientifica?» *Ecdotica* 1 (2019): 174–85.

[12] Pierazzo, Elena. «What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter». *International Journal of Digital Humanities* 1, fasc. 2 (2019): 209–20.

[13] Ruffini, Elisabetta. «Come se scoprissero una popolazione sconosciuta della Nuova Guinea". L'immediato dopoguerra e i primi libri sul Lager». In *Riga 38. Primo Levi*, (a cura di) Mario Barenghi, Marco Belpoliti, e Anna Stefi, 557–63. Milano: Marcos y Marcos, 2017.

[14] Samuel, Jean, e Jean Marc Dreyfus. *Il m'appelait Pikolo. Un compagnon de Primo Levi raconte*. Tradotto da Claudia Lionetti Frassinelli. Paris: Laffont, 2007.

[15] Schmidt, Desmond. «True interoperability for digital scholarly editions». *Umanistica Digitale* 10 (2021): 1–23.

[16] Tesio, Giovanni. «Su alcune giunte e varianti a *Se questo è un uomo*». *Studi Piemontesi* VI, fasc. 2 (1977): 270-78.

Per un'edizione scientifica digitale dello *Speculum Guy of Warwick*

Omar Khalaf[1], Sibilla Siano[2]
[1] Università degli Studi di Padova, Italia – omar.khalaf@unipd.it
[2] Università degli Studi di Padova, Italia – sibilla.siano@unipd.it

ABSTRACT
All'inizio del quattordicesimo secolo, nello stesso momento in cui il romanzo *Guy of Warwick* è attestato per la prima volta in medio inglese, la tradizione della Materia di Inghilterra sembra essersi arricchita di un altro testo intorno a questo eroe: il poema didattico devozionale *Speculum Guy of Warwick* (*SGW*). I suoi 1034 versi in distici rimati sono interamente dedicati all'istruzione dell'eroe nazionale ad opera del sapiente Alcuino di York. Nonostante la fortuna di cui questo poema ha goduto in epoca medievale, *SGW* sembra essere stato quasi completamente ignorato in epoca più recente. Tuttavia, la ricchezza della sua tradizione testuale merita di essere studiata e approfondita, in primo luogo attraverso un'edizione che prenda in considerazione la totalità dei testimoni pervenuti e che dia conto della varietà che contraddistingue le varie redazioni. Il progetto qui presentato si propone di realizzare un'edizione scientifica digitale di *SGW* attraverso la quale al testo criticamente stabilito possa essere accostato uno strumento in grado di valorizzare le specificità di ogni singolo testimone.

PAROLE CHIAVE
Speculum Guy of Warwick; Middle English; SDE; critical edition; EVT 2.

1. INTRODUZIONE
Il testo conosciuto come *Speculum Guy of Warwick* (*SGW*) è un poema didattico devozionale in medio inglese che ha per oggetto l'istruzione dell'eroe d'Inghilterra Guy of Warwick ad opera del sapiente Alcuino di York. Nella sua forma completa, occupa poco più di mille versi in distici rimati. Mentre il romanzo dedicato a Guy fu composto inizialmente in anglo-normanno per poi essere tradotto nell'altra lingua vernacolare d'Inghilterra, *SGW* non sembra avere antecedenti in lingue altre. La fortuna di quest'opera è dimostrata dalla ricchezza della sua tradizione testuale, che conta dieci testimoni pervenuti sino a noi [5][1]. La versione più antica risale agli anni Trenta del quattordicesimo secolo ed è contenuta nel celebre manoscritto Auchinleck (**A1**), nel quale co-occorre con i romanzi dedicati a Guy of Warwick. È forse proprio in questo contesto che l'eroe nazionale subisce la definitiva trasformazione in *miles Christi*, passando da una caratterizzazione tipica della tradizione cavalleresca ad una di stampo puramente religioso [5: 82]; negli altri testimoni è infatti spesso accompagnato da numerosi testi devozionali [1: 11, 2: 81. 6: 275], a riprova del fatto che il legame con i romanzi che avevano dato fama all'eroe eponimo andava via via affievolendosi.
Nonostante l'incredibile ricchezza di riferimenti intertestuali e l'affascinante relazione con la più studiata figura eroica del protagonista, questo poema ha ricevuto un'attenzione accademica decisamente modesta. L'unica edizione disponibile risale al 1898 e si basa su sei dei dieci manoscritti sopra menzionati. Il ritrovamento di nuovi codici successivamente alla pubblicazione dell'edizione di Morrill [6] non può che mettere in discussione lo *stemma codicum* ipotizzato. Ad esempio, secondo quanto evidenziato dalla studiosa, **A1** sarebbe il più antico testimone di *SGW*, seguito a distanza di mezzo secolo da **R**. Tuttavia, la successiva scoperta di un'ulteriore redazione del poema essenzialmente coeva ad **A1** nel manoscritto **C** potrebbe alterare significativamente le ipotesi fin qui proposte sulla tradizione testuale dell'opera. Un' analisi comparativa della versione di **A1** con quella di **C** potrebbe infatti far luce sulle prime fasi della tradizione del testo. Inoltre, le diverse edazioni di *SGW* presentano un livello di variazione testuale che non è mai stato dovutamente considerato; esso, tuttavia, dimostra una ricezione dell'opera tutt'altro che passiva e fornisce la cifra del suo livello di adattabilità tematica e contenutistica nel tempo e nello spazio.

[1] Edinburgh, National Library of Scotland MS Advocates' 19.2.1 [**A1**]; London, British Library MS Arundel 140 [**A2**]; London, British Library MS Add. 36983 [**A3**]; Cambridge, St John's College MS S.30 (256) [**C**]; Cambridge, Cambridge University Library MS Dd. 11.89 [**D**]; London, British Library MS Harley 1731 [**H1**]; London, British Library MS Harley 525 [**H2**]; Manchester, John Rylands Library MS Eng. 50 [**M**]; Oxford, Bodleian Library MS Add. C. 220 [**O**]; London, British Library MS Royal 17 B.XVII [**R**]. Si veda anche il *Digital Index of Middle English Verse* (DIMEV), https://www.dimev.net/record.php?recID=1782. Come **C**, anche **A3** è stato scoperto successivamente alla pubblicazione dell'edizione di Morrill [6]; non essendo ancora il suo testo nella nostra disponibilità, è escluso dalla presente disamina.

2. LA TRADIZIONE TESTUALE DI SGW: *VARIATIO* E *INNOVATIO*

La ricezione attiva di *SGW* nel corso del tempo è ampiamente dimostrata dalla *mouvance* [7] che caratterizza la sua trasmissione testuale e che suggerisce, come nel succitato caso di **A1**, che il testo sia stato recepito e rimodellato dai compilatori dei manoscritti per soddisfare esigenze specifiche a livello di interpretazione e di funzione di ogni testimone nel suo contesto codicologico. I vari testimoni si caratterizzano per una marcata variazione lessicale, ma numericamente rilevanti sono anche i casi di tagli e interpolazioni di interi versi, i quali modificano sensibilmente il contenuto di ogni singola redazione. Gli studi condotti sull'opera finora, a partire dall'edizione di Morrill, non hanno dovutamente messo in luce questo aspetto, che purtuttavia si rivela fondamentale per uno studio della ricezione di questo testo. L'editrice basa la sua edizione su **A1**, un testimone reso incompleto dal danneggiamento dell'ultimo foglio e la cui parte mancante è stata ricostruita sulla base di **D**; il suo testo, quindi, si configura come una sorta di ibrido in cui la redazione imperfetta di **A1**, scelta probabilmente in ossequio all'antichità del manoscritto, viene integrata con un secondo testimone la cui scelta non è però stata giustificata dalla studiosa. Dal punto di vista filologico questa prassi non trova motivazioni evidenti: l'archetipo di *SGW* poteva essere individuato in **R** (ora, sappiamo, anche in **C**), in quanto preserva il testo completo. Le varianti presenti negli altri quattro manoscritti considerati (**A2**, **H1**, **H2** e **R**) sono relegate indiscriminatamente all'apparato critico, con il risultato che la ricchezza della tradizione del testo così come la specificità di ogni singolo testimone risultano marginali e appiattite nell'intento ricostruttivo tipico delle edizioni tradizionali. Ad esempio, dopo il v. 4 il solo **H2** presenta un'interpolazione di tre versi:

For the sowlys saluacyowne
Who soo that herythe þis sermoune
Inicium sapiencie timor Domini.

Il redattore di questo testimone trasforma *Inicium sapiencie timor Domini*, che tutte le altre redazioni riportano come il titolo della sezione finale del sermone di Alcuino (v. 883), come il nome dell'intero sermone, operando quindi una profonda modificazione della struttura del testo. Morrill accoglie questa e tutte le altre variazioni nell'apparato critico, privando **H2**, così come l'intera tradizione, della profondità storica che caratterizza la trasmissione di quest'opera.

Pertanto, nello specifico di *SGW* i limiti metodologici e pratici di un lavoro ecdotico fondato su basi ricostruttive non rendono giustizia alla ricchezza della tradizione testuale dell'opera e alla sua ricezione nel tempo e nello spazio. In simili casi di recensione aperta [4], la ricostruzione di un archetipo, per quanto utile a offrire un testo stabilito e fruibile al pubblico moderno, limita gravemente l'individuazione e l'analisi delle caratteristiche proprie di ogni singolo testimone; d'altra parte, il numero non irrilevante di testimoni renderebbe assai difficilmente praticabile la realizzazione di un'edizione sinottica secondo gli standard editoriali propri delle edizioni cartacee.

Una tradizione così ricca di variazioni e innovazioni, perciò, merita di essere valorizzata attraverso uno strumento che permetta di rendere conto di tale complessità testuale. A tal fine, il presente progetto di ricerca ambisce a realizzare un'edizione che offra al pubblico un testo criticamente stabilito di *SGW* attraverso la collazione di tutti i testimoni pervenuti e, allo stesso tempo, permetta di restituire la profondità della sua tradizione manoscritta. Il fine ultimo di questo prodotto è di fornire uno strumento che possa favorire ulteriori indagini future sulla complessità che caratterizza questo testo, così significativo ma allo stesso tempo così poco studiato finora; tutte queste caratteristiche troveranno pertanto applicazione nell'edizione scientifica digitale di *SGW*, nelle modalità specificate di seguito.

3. PER UN'EDIZIONE SCIENTIFICA DIGITALE DI *SGW*

In quest'ottica, il gruppo di ricerca si propone di realizzare di un'edizione che possa rendere conto di tale variazione a livello intrastemmatico e, allo stesso tempo, fornire all'utente uno strumento in grado di raccogliere e sistematizzare, secondo livelli ecdotici specifici, la *variatio* che caratterizza *SGW*, le sue modalità di trasmissione e la sua ricezione per come espresso dalla tradizione testuale a noi pervenuta. A tal fine, lo strumento digitale si rivela essere particolarmente efficace nella sua capacità di offrire, in un contesto interattivo e personalizzabile, un prodotto che permetta la visualizzazione dell'edizione critica di *SGW* e un confronto attivo tra i vari testimoni.

Per questo motivo si è deciso di progettare un'edizione scientifica digitale di *SGW* e di affidarsi al software *Edition Visualization Technology* (EVT 2)[2], sfruttando la sua versatilità e le molteplici possibilità che offre. Nella sua *release* attuale [5] il software costituisce lo strumento ideale ai fini del progetto, soprattutto per quanto riguarda: a) la

[2] http://evt.labcd.unipi.it/

corrispondenza delle sue funzionalità agli obiettivi dell'edizione progettata; b) la sua adesione a protocolli di codifica standard; c) la facilità di utilizzo e personalizzazione; d) la piena rispondenza del prodotto finale ai principi FAIR.
Dal punto di vista operativo, EVT 2 soddisfa in pieno le esigenze ecdotiche specifiche per il presente progetto: il livello critico, costruito in modo da consentire un alto livello di interattività, supporta un apparato positivo nel quale l'utente ha la possibilità di cliccare sul *siglum* di qualsiasi altro testimone ed accedere non solo al testo completo di quest'ultimo, ma anche visualizzare note, commenti o rimandi a fonti. Inoltre, l'attivazione della modalità "collazione" permette una visualizzazione sinottica dei testimoni; l'evidenziazione delle varianti presenti nello stesso luogo testuale al passaggio del cursore permette un confronto testuale immediato, impossibile da operare in un'edizione tradizionale cartacea. Tali funzioni si rivelano pienamente rispondenti alle esigenze di un'edizione di *SGW* progettata per rendere conto della sua spiccata mobilità testuale. Una panoramica sul *workflow* della codifica e dell'allestimento del software è fornita nella sezione seguente.

4. CODIFICA E VISUALIZZAZIONE

Di seguito si illustreranno gli aspetti metodologici e applicativi salienti relativi all'attività di marcatura del testo e del suo riversamento in EVT 2. Per la creazione dell'apparato critico, il software supporta la modalità "Parallel segmentation" come dal protocollo TEI-P5: essa prevede l'utilizzo dell'elemento `<app>`, il quale codifica la lezione ritenuta corretta (codificata attraverso `<lem>`) e le varianti presenti negli altri testimoni (marcati con `<rdg>`)[3]. Pertanto, la prassi segue fedelmente il criterio lachmanniano nella scelta della lezione più accettabile rispetto al resto della tradizione, aderendo, per la maggior parte dei casi, al principio dell'*ope codicum*. Un esempio è dato dal primo verso del poema, che nella sua forma ricostruita riporta "Herkeneþ alle to my speche" ("Ascoltate tutti il mio racconto"); solo **C** si discosta dal resto della tradizione in quanto preserva "Lestnet alle to my spelle". La scelta di codificare "herkeneþ" e "speche" con `<lem>` e "lestneth" e "spelle" con `<rdg>` si fonda su un ovvio parametro statistico, giacché la maggioranza dei testimoni riporta le prime due lezioni; lo stesso può dirsi della preposizione "to", attestata in tutti i testimoni ad eccezione di **D** e **H2**, i quali tramandano "vnto"; la marcatura del verso, di conseguenza, segue la struttura seguente (vd. Fig.1):

```
<l>
        <app>
            <lem wit="#A1 #A2 #D #H1 #H2 #O #R">Herkne&#254;</lem>
            <rdg wit="#C">Lestnet</rdg>
            <rdg wit="#M">Herknes</rdg>
        </app>
        alle
        <app>
            <lem wit="#A1 #A2 #C #H1 #M #O #R">to</lem>
            <rdg wit="#D #H2">vnto</rdg>
        </app>
        my
        <app>
            <lem wit="#A1 #A2 #D #H1 #H2 #M #O #R">speche</lem>
            <rdg wit="#C">spelle</rdg>
        </app>
        </l>
        <l>
```

Figura 1. Codifica del verso 1

Ove opportuno, al criterio meccanico si affiancheranno interventi *ope ingenii*. Al verso 23, che nell'edizione riporta "For swiche þer beþ þat loueþ more" ("Perché ce ne sono [di uomini] tali che amano di più"), si è preferito l'avverbio *þer*, attestato in **H1**, **H2** e **O** in luogo del pronome *it*, preservato in un numero maggiore di testimoni (**A1**, **D** e **R**, con la variante hit in **C** e **M**); in questo caso, la scelta è stata operata applicando criteri squisitamente linguistici, che si sono imposti sul criterio statistico (vd. Fig.2).

[3] "12.2.3. The Parallel Segmentation Method", in *TEI: Guidelines for Electronic Text Encoding and Interchange P5 Version 4.7.0*, disponibile al link https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS.

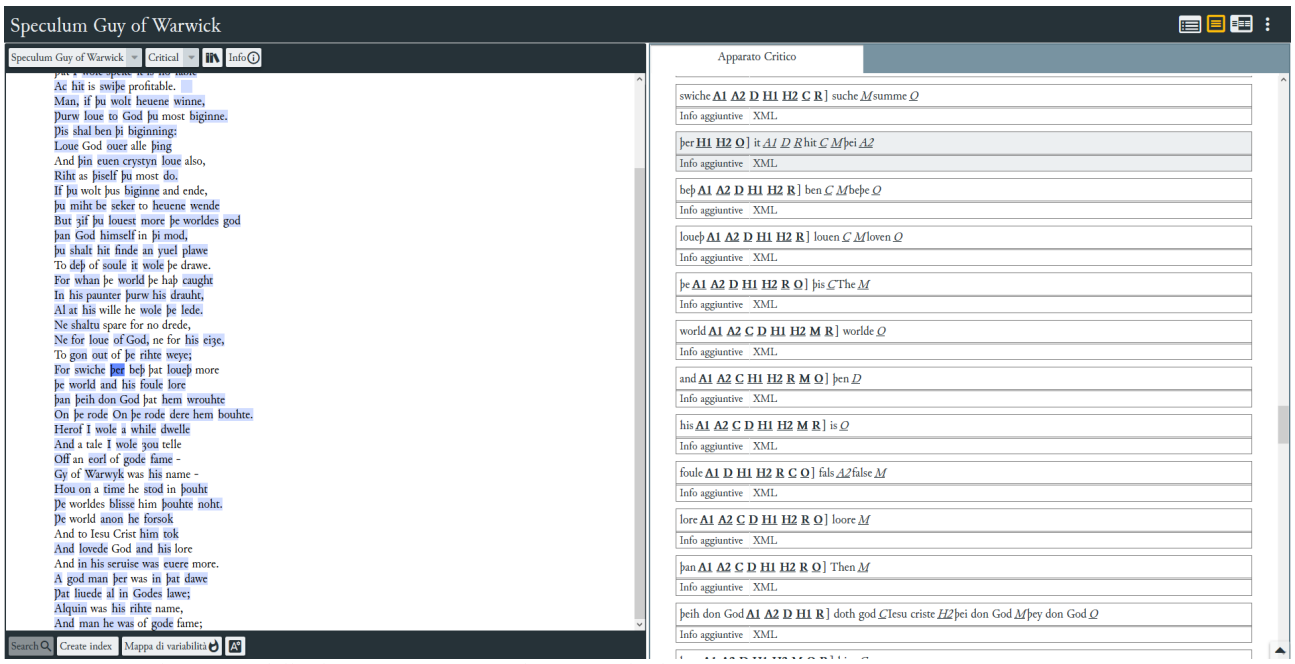


Figura 2. Visualizzazione dell'edizione in EVT 2 con l'esempio di "þer"

Come evidenziato nella Fig. 3, il riversamento della codifica in EVT 2 permette una visualizzazione soddisfacente dell'edizione così prodotta; al testo stabilito si affianca l'apparato critico positivo, il quale distingue la lezione scelta rispetto alle varianti attestate nel resto della tradizione attraverso l'evidenziazione in grassetto dei testimoni che la contengono, così da permetterne una loro immediata identificazione.
Il software supporta efficacemente anche la visualizzazione in apparato della summenzionata interpolazione in **H2** tra i versi 4 e 5; l'inserzione dell'elemento `<nota>` nel nodo `<app>` in fase di codifica, inoltre, permette di corredare l'apparato con informazioni utili alla sua interpretazione.

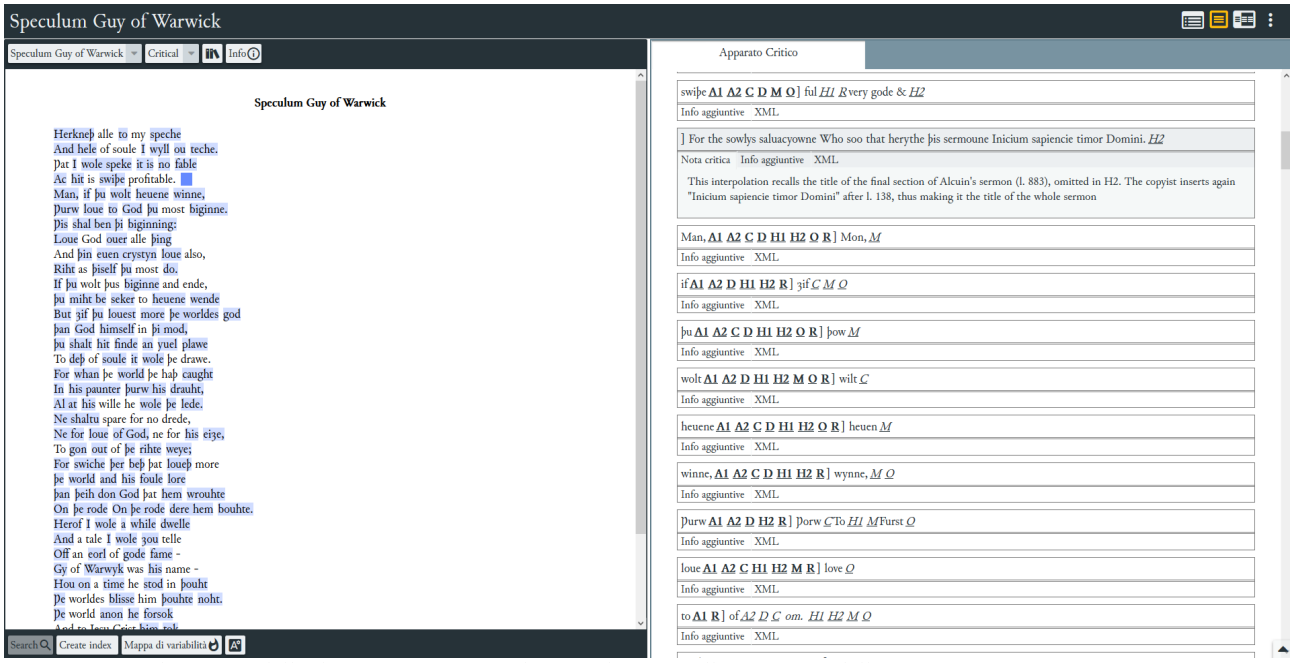


Figura 3. Visualizzazione dell'edizione in EVT 2 con l'interpolazione nell'apparato e della nota critica

L'aggiunta trova la sua corretta collocazione testuale in **H2** grazie all'attivazione della modalità "collazione" prevista in EVT 2 (vd. Fig.4). Di fronte ad una tradizione variabile come quella di *SGW*, la possibilità di fornire all'utente l'accesso

alla versione del testo come preservata in un singolo testimone e di accostarla sinotticamente a quella attestata in qualsiasi altro manoscritto[4] costituisce un vantaggio che le edizioni cartacee non possono soddisfare.

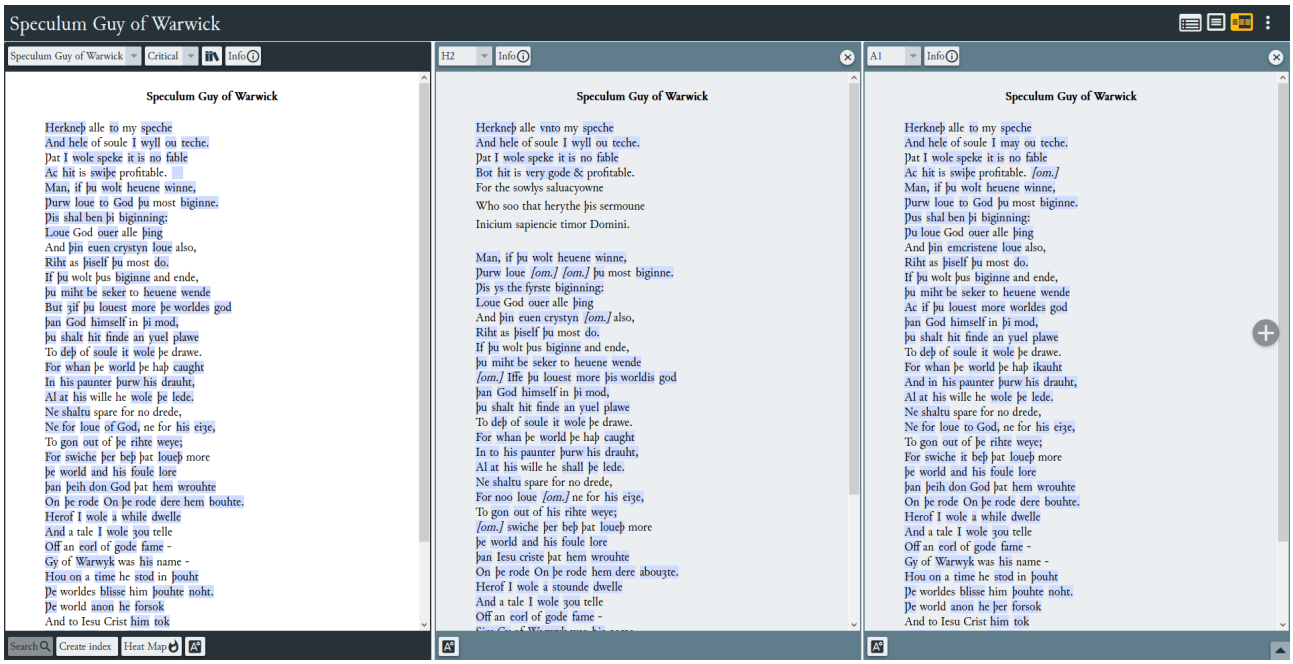


Figura 4. Modalità "collazione", attraverso la quale è possibile, ad esempio, confrontare il testo edito, H2 e A1

Di default il software prevede una sezione "Info", posizionata nell'area superiore della pagina. Questa funzione è stata implementata con la redazione di un breve paragrafo con cenni sulla storia di *SGW* e sullo *status quaestionis* e, a seguire, l'elenco dei testimoni con informazioni codicologiche di base (vd. Fig. 5). La parte verrà progressivamente ampliata, anche in virtù di future indagini da parte del gruppo di ricerca.

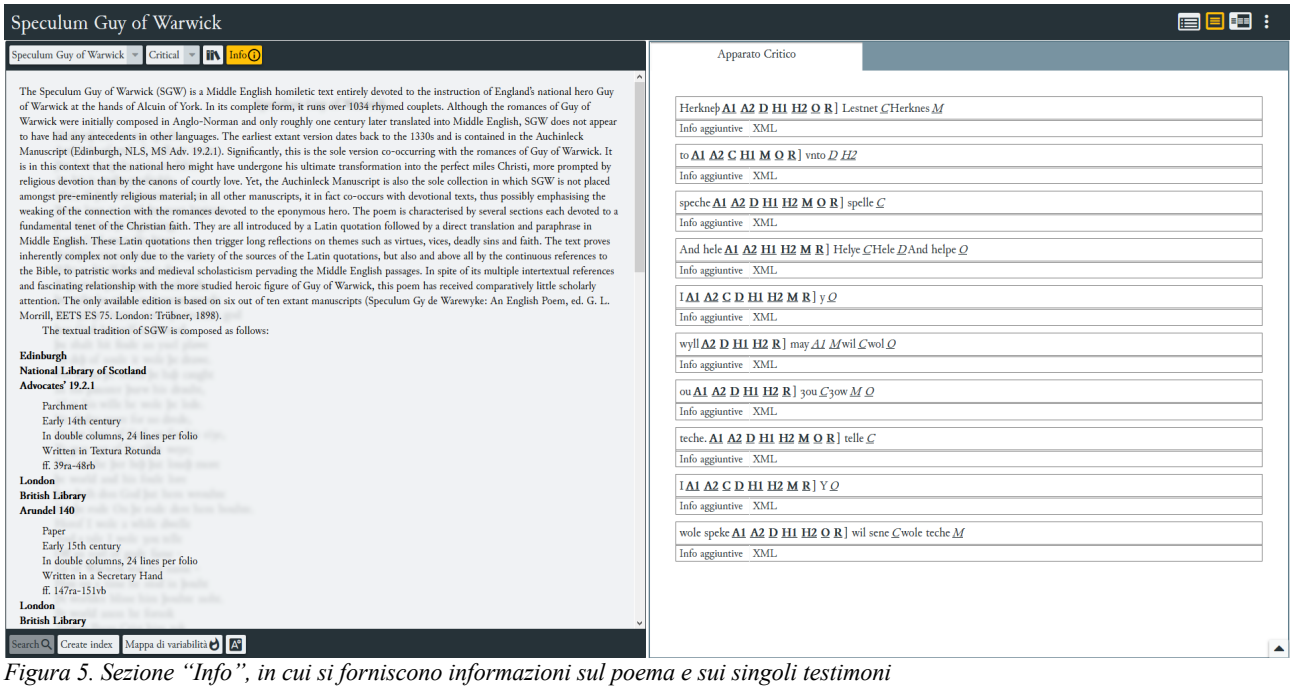


Figura 5. Sezione "Info", in cui si forniscono informazioni sul poema e sui singoli testimoni

Allo stato attuale il progetto è ancora nelle sue fasi preliminari; a fronte della disponibilità delle trascrizioni di tutti i testimoni, la collazione di tutto il testo non è ancora stata completata e pertanto non si esclude di dover implementare, per

[4] Per uno schermo di dimensioni standard, la modalità ottimale prevede la visualizzazione contemporanea del testo critico accanto a due testimoni.

casi specifici che dovessero presentarsi, la progettazione della codifica. Alcune problematiche di carattere strutturale dovranno essere risolte, come, ad esempio, la sfasatura nell'indicazione dei numeri di verso tra l'edizione critica e i singoli testimoni nella modalità "collazione", a causa della quale si è deciso, almeno al momento, di non inserire alcun riferimento, o questioni di spaziatura che incidono sul layout del testo di **H2** in corrispondenza dell'interpolazione (vd. Fig. 4); un altro problema riguarda la visualizzazione della punteggiatura, la quale, in caso di occorrenza dopo una parola oggetto di intervento critico, è stata inserita all'interno dell'elemento `<app>` per evitare una spaziatura indesiderata tra quest'ultima e il segno d'interpunzione. Inoltre, ci si proporrà di lavorare sulla configurazione di EVT 2 per permettere, sempre nella modalità "collazione", lo scorrimento parallelo di tutti i testimoni visualizzati, così da permettere una più agevole comparazione.

Tuttavia, la progettazione della *SDE* così com'è stata qui brevemente illustrata sembra rispondere pienamente agli obiettivi che il gruppo di ricerca sta perseguendo: fornire uno strumento utile all'indagine di un testo dimenticato per decenni dalla critica, ma che ha goduto di una grandissima fortuna nell'Inghilterra medievale e, per questo motivo, meritevole di essere apprezzato e studiato.

5. RINGRAZIAMENTI

Il progetto è finanziato dal programma di ricerca PRIN 2022SEK27B, *Politics of Worship pre- and post-Reformation*, di cui il Dipartimento di Studi Linguistici e Letterari dell'Università degli Studi di Padova è partner.

BIBLIOGRAFIA

[1] Edwards, Anthony S.G. «The Speculum Guy de Warwick and Lydgate's Guy of Warwick: The Non-Romance Middle English Tradition». In *Guy of Warwick: Icon and Ancestor*, a cura di Alison Wiggins e Rosalind Field, 81–93. Woodbridge: D. S. Brewer, 2007.

[2] Hume, Cathy. *Middle English Biblical Poetry: Romance, Audience and Tradition*. Woodbridge: D. S. Brewer, 2021.

[3] Murchinson, Krista A. *Manuals for Penitents in Medieval England*. Woodbridge: D. S. Brewer, 2021.

[4] Pasquali, Giorgio. *Storia Della Tradizione e Critica Del Testo*. Vol. 1. Firenze: Le Monnier, 1934.

[5] Rosselli Del Turco, Roberto, Chiara Di Pietro, e Chiara Martignano. «Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo». *Umanistica Digitale*, 2019, No 7 (2019). https://doi.org/10.6092/ISSN.2532-8816/9322.

[6] Warewyke, Speculum Gy de. *An English Poem*. A cura di Georgiana L. Morril. EETS ES 75. London: Trübner, 1898.

[7] Zumthor, Paul. *Essai de poétique médiévale*. Paris: Seouil, 1972.

Per una lettura antropologica di Verga: tra codifica e georeferenziazione

Giovanna Zisa
Università di Catania, Italia - giovanna.zisa@phd.unict.it

ABSTRACT

Il contributo intende presentare una sezione del progetto di ricerca dal titolo *Giovanni Verga tra demopsicologia e antropologia. Mappatura GIS e database delle attestazioni folkloriche nel corpus verghiano: creazione di uno spazio digitale*. Il progetto è ancora in una fase iniziale; qui ci si propone di illustrarne le principali linee di sviluppo e i punti di contatto con un progetto più ampio, PNRR - "CHANGES" SPOKE 3, che prevede la realizzazione di edizioni digitali commentate di testi della letteratura italiana del XIX secolo e, nello specifico, dei tre grandi autori del Verismo: Luigi Capuana, Federico De Roberto e Giovanni Verga.

Dopo aver esposto il progetto nella sua totalità ci si concentrerà su una sezione di esso, descrivendo, anche attraverso la presentazione di un esempio concreto tratto da *Vita dei campi*, il "*database* folklorico*"* che si vuole realizzare e mostrando i risultati di una prima georeferenziazione dei luoghi delle novelle *Fantasticheria* e *Cavalleria rusticana*.

Si illustreranno le fasi di modellizzazione del *database* e, dato che sia per la creazione di quest'ultimo sia per la realizzazione dell'edizione digitale di *Vita dei campi* ci si servirà della codifica in XML/TEI, parte del contributo sarà dedicata a quest'ultima, al fine di esplicitarne i vantaggi e le potenzialità nell'ambito della gestione, dell'archiviazione e dell'implementazione dei dati.

L'impatto atteso sarà valutato da due prospettive differenti: da un lato si considereranno i vantaggi di questa operazione ai fini di uno studio tradizionale dell'autore, dall'altro si esamineranno le agevolazioni alla ricerca fornite dalla creazione di uno spazio digitale aperto e condiviso.

PAROLE CHIAVE

Coding; XML/TEI; Giovanni Verga; demopsychology; modeling.

1. INTRODUZIONE: LINEE GUIDA DEL PROGETTO

La prima fase del lavoro si inserisce all'interno del progetto PNRR - "CHANGES" SPOKE 3 e, nello specifico, nella sezione *Verismo digitale* che ha l'obiettivo di realizzare le edizioni digitali commentate delle seguenti opere di Giovanni Verga: *Vita dei campi*, *I Malavoglia*, *Novelle rusticane* e *Mastro-don Gesualdo*. Per quanto concerne *Vita dei campi*, il testo su cui si è scelto di eseguire la codifica in XML/TEI è quello della *princeps*, già fissato come testo di riferimento da Carla Riccardi che ne ha curato l'Edizione Nazionale nel 1987.

La seconda fase del progetto prevede la realizzazione di un commento dell'opera che faccia riferimento alla critica verghiana del passato e che integri a essa gli studi più recenti, per fornire al lettore un quadro completo e per giungere a un'interpretazione quanto più possibile accurata.

L'ultima fase del progetto si propone di rintracciare nel *corpus* verghiano, e in particolar modo in *Vita dei campi*, opera che meglio delle altre si è prestata e ancora oggi si presta a una lettura di tipo antropologico, riferimenti al folklore che permettano di istituire un confronto tra Verga e i demologi siciliani dell'Ottocento. L'intento profondo è di analizzare gli influssi del nuovo positivismo antropologico europeo sulle opere di Verga e di cogliere analogie e differenze tra l'autore e gli etnologi e folkloristi siciliani. Nell'ambito di tale disamina verranno indagate le opere di Giuseppe Pitrè, Serafino Amabile Guastella, Santo Rapisarda, Lionardo Vigo Calanna, Salvatore Salomone Marino e si approfondirà lo studio delle riviste di demopsicologia dell'epoca. L'indagine si concentrerà inoltre sulla mappatura dei luoghi del *corpus* verghiano e delle opere demologiche di interesse: questo tipo di operazione, oltre ad agevolare la comparazione tra gli autori, servirà a ribadire il legame profondo tra folklore e territorio.

2. RIFERIMENTI TEORICI E PRESUPPOSTI DI RICERCA

Gli strumenti digitali per l'archiviazione offrono oggi molte opportunità all'umanista che vuole organizzare e gestire una grande mole di dati. Il concetto di *database* non è di recentissima data, ma gli studi sulla funzione delle basi di dati in ambito umanistico sono in continuo aggiornamento e costituiscono una parte fondamentale della disciplina Informatica Umanistica. La progettazione di un *database* implica una riflessione più ampia sul significato e sull'importanza della modellizzazione e, più in generale, sulla pratica del *Project Management.* Quest'ultima, come evidenziato da Daniele

Marotta, spesso non viene integrata «[…] nella definizione di modelli, metodologie e strumenti utilizzati nelle ricerche umanistiche» [10: 89] poiché se ne sottovalutano le potenzialità e i vantaggi, specialmente in termini di "tracciabilità" e di replicabilità del lavoro, concetti che Cristina Marras chiarisce, attraverso la metafora dell'ecosistema, parlando di "riciclo": «un ecosistema si considera sostenibile se ha la possibilità, anche, di riciclarsi» [11]. Affinché un progetto sia "riciclabile", e quindi "sostenibile", non basta però rendere tracciabili tutte le sue attività: l'utilizzo della modalità *open source* e l'adozione di *standard* condivisi sono essenziali ai fini di garantire l'interoperabilità e la condivisione dei dati e per scongiurarne l'obsolescenza. È interessante ricordare, inoltre, che di recente il concetto di "*Open Access*" è stato significativamente legato a quello di "cittadinanza scientifica"; come afferma Paola Castellucci, il periodo di reclusione causato dalla pandemia ha reso ancora più urgente una profonda riflessione sull'*Open Science*: «La cultura open è parsa una precondizione per un discorso consapevole rispetto al diritto all'accesso all'informazione e alla conoscenza; una risposta armonica, politica e scientifica, rispetto ai bisogni profondi della società […]» [3: 225].

Tornando brevemente alla modellizzazione, Arianna Ciula e Cristina Marras sostengono che «La modellizzazione in Digital Humanities esplicita le componenti umanistiche e computazionali mettendole in relazione» [5: 64]. In questo senso modellizzare significherebbe astrarre gli "oggetti culturali" dalla loro materialità (di cui bisogna comunque tenere conto per modellizzare in modo adatto) per inserirli in un contesto computazionale più ampio che permetta di creare delle relazioni complesse tra dati resi "misurabili" (e quindi "processabili") dal mezzo digitale. L'importanza della modellizzazione è ancora più chiara se la si pensa in relazione ai *database*; come spiega Francesca Tomasi, già durante la fase di progettazione concettuale, attraverso la creazione di un modello si può avere «[…] una rappresentazione formale del *corpus* in questione, indipendentemente dalla concreta realizzazione del DB e dell'ambiente tecnologico che si intende utilizzare» [15: 89]. Avere un modello che consenta di visualizzare le relazioni tra i dati significa chiarire a che fine si intendono raccogliere questi ultimi e poter illustrare lo scopo della ricerca con più consapevolezza.

In ambito umanistico i dati con cui si ha più spesso a che fare sono quelli testuali; per modellizzarli, già a partire dagli anni '90, è stata utilizzata prevalentemente la codifica in XML/TEI, la quale ha permesso di «[…] definire uno standard comune nella comunità scientifica degli "umanisti digitali" per la rappresentazione di testi su supporto informatico» [4: 88], rendendo i dati interoperabili (e, dunque, "riutilizzabili") e agevolando quell'interscambio oggi tanto auspicato e necessario per poter parlare concretamente di *cultura open*. Mettere in rete edizioni digitali attendibili in modalità *open access* [6] significa sancire «[…] il diritto alla "cittadinanza scientifica" per tutti, e non solo per chi fa ricerca, o studia, per professione» [3: 216-217] ma significa anche arginare la proliferazione di edizioni digitali (o di digitalizzazioni) che, sebbene utili alla divulgazione, spesso non sono filologicamente attendibili [9: 208]. Fornire edizioni filologicamente accurate, mettere a disposizione dell'utente un commento su più livelli che agevoli e chiarifichi la lettura dell'opera, è un'operazione particolarmente importante soprattutto se attuata su quei testi che fanno imprescindibilmente parte dei programmi scolastici e che sono letti e conosciuti su larga scala. Per quanto concerne, nello specifico, Giovanni Verga, mi sembra che l'avanzamento scientifico che si è registrato sul piano critico non sia stato accompagnato da un adeguato progresso in senso tecnologico e multimediale; gli studi che mirano ad approfondire il rapporto tra Verga e il folklore, ad esempio, di recente tornati alla ribalta grazie alle indagini di specialisti in scienze demo-etnoantropologiche come Lia Giancristofaro e Mauro Geraci, e grazie a studiosi di Letteratura italiana antropologica come Riccardo Castellana, hanno riportato l'attenzione sugli studi di demopsicologia ottocenteschi e sulla loro influenza nell'opera verghiana [2, 8, 7], ma questo interesse non è stato accompagnato da una riflessione sulle possibilità offerte dal mezzo digitale nella sfera della visualizzazione dei risultati della ricerca, anche in prospettiva di una maggiore condivisione.

A proposito di resa visuale dei dati e di condivisione degli stessi, la geografia offre un supporto a discipline che, come la letteratura, sono eminentemente teoriche, favorendo oggi, anche grazie all'ausilio di strumenti tecnologici più all'avanguardia, lo sviluppo di analisi interdisciplinari che coniugano saperi diversi al fine di realizzare prodotti innovativi e interattivi [13: 3]. I Sistemi Informativi Geografici (GIS), ad esempio, permettono di svolgere una serie di operazioni che consentono allo studioso di organizzare e diffondere la conoscenza in modo accattivante e di associare a un'analisi di tipo economico-sociale quella ambientale. La creazione di *geodatabase* in ambiente GIS, ad esempio, dà la possibilità di raggruppare dati provenienti da fonti diverse e «[…] aiuta a porre le fondamenta per restituire – secondo opportune metodologie e funzionalità applicative – cartografie digitali e prodotti multimediali funzionali ad analisi puntuali e alla veicolazione dei risultati raggiunti» [13: 4]. Poter creare *database* in cui al dato geografico si associ, ad esempio, quello bibliografico, è per lo studioso di discipline umanistiche un'opzione ricca di potenzialità, resa ancor più interessante dalla possibilità di caricare sul Web mappe dinamiche che integrino e mostrino le informazioni su diversi livelli [14: 258].

3. METODOLOGIA: LA MODELLIZZAZIONE DEL DATABASE

Per la realizzazione delle edizioni digitali previste dal progetto PNRR - "CHANGES" SPOKE 3 sono state redatte delle indicazioni di marcatura comuni: dopo aver inserito gli elementi di segmentazione testuale (`<div>`, `<p>`) si è deciso di marcare persone, luoghi, oggetti, elementi lessicali, organizzazioni e date e di segnalare la presenza del discorso diretto attraverso l'utilizzo del tag `<q>`. Dell'opera *Vita dei campi,* al momento, è stata realizzata per intero la codifica strutturale ed è in corso la marcatura dei nomi di persona e di quelli di luogo, già completata nelle novelle *Fantasticheria* e *Cavalleria rusticana* (vd. Fig. 1).

```
1103    <div type="novella">
1104    <pb n="75"/>
1105    <head>Cavalleria rusticana</head>
1106    <p><persName ref="TuridduMacca" type="fictional">Turiddu Macca</persName>, il figlio della
1107    gnà <persName ref="Nunzia" type="fictional">Nunzia</persName>, come tornò da fare il soldato, ogni domenica si pavoneggiava
1108    in piazza coll'uniforme da bersagliere e il berretto rosso, che sembrava quello della buona ventura, quando mette su banco
1109    colla gabbia dei canarini. Le ragazze se lo rubavano cogli occhi, mentre andavano a messa col naso dentro la mantellina, e i
1110    monelli gli ronzavano attorno come le mosche. Egli aveva portato anche una pipa col re a cavallo che pareva vivo, e accendeva
1111    gli zolfanelli sul dietro dei calzoni, levando la gamba, come se desse una pedata. Ma con tutto ciò
1112    <persName ref="Lola" type="fictional">Lola</persName> di massaro <persName ref="Angelo" type="fictional">Angelo</persName>
1113    non si era fatta vedere né alla messa, né sul ballatoio ché si era fatta sposa con uno di
1114    <placeName ref="Licodia" type="real">Licodia</placeName>, il quale faceva il carrettiere e aveva quattro muli di
1115    <placeName ref="Sortino" type="real">Sortino</placeName> in stalla.
```

Figura 1

Per indagare i dati antropologici all'interno del *corpus* verghiano e rendere visualizzabili le corrispondenze con le opere dei demologi, si è scelto di sfruttare le possibilità offerte, nella codifica XML/TEI, dagli attributi `@source` e `@xml:id`. L'esempio che segue è tratto da *Vita dei campi* ma, poiché nell'ambito del progetto *Verismo digitale* saranno codificate anche le altre opere maggiori di Verga, i riferimenti folklorici potranno essere ricercati in un *corpus* più ampio e articolato e, dunque, più significativo anche dal punto di vista dei risultati della ricerca. Le opere dei demologi non saranno codificate per intero: i segmenti testuali di riferimento saranno estrapolati e marcati in un documento denominato "Demologi.xml".
L'organizzazione dei dati nel documento Vita_dei_campi.xml permetterà il collegamento tra i diversi documenti e avverrà come segue:

1) il tag `<seg>` delimiterà la porzione di testo oggetto di indagine;
2) attraverso l'attributo `@type` si assegnerà al segmento testuale selezionato il valore "riferimento_folklorico";
3) l'attributo `@subtype` specificherà ulteriormente il valore del "riferimento_folklorico" in questione, descrivendolo come proverbio, festa, oggetto folklorico, ecc.
4) Alla parte di testo descritta come proverbio, festa, oggetto folklorico, sarà assegnato un ulteriore attributo `@source` che permetterà il collegamento con il file Demologi.xml e, nello specifico, con il segmento di testo oggetto di comparazione.

Si riporta di seguito un esempio tratto dalla novella *Cavalleria rusticana* dalla quale si è selezionato come riferimento folklorico il proverbio "facemu cuntu ca chioppi e scampau, e la nostra amicizia finiu" (vd. Fig. 2).

```
<p><q>– È giusto</q>, rispose <persName ref="TuridduMacca" type="fictional">Turiddu</persName>; <q>ora che sposate
compare <persName ref="Alfio" type="fictional">Alfio</persName>, che ci ha quattro muli in stalla, non bisogna farla chiacchierare la gente. Mia madre invece,
poveretta, la dovette vendere la nostra mula baia, e quel pezzetto di vigna sullo stradone, nel tempo ch'ero soldato. Passò quel tempo che Berta filava, e voi non
ci pensate più al tempo in cui ci parlavamo dalla finestra sul cortile, e mi regalaste quel fazzoletto, prima d'andarmene, che Dio sa quante lagrime ci ho pianto
dentro nell'andar via lontano tanto che<pb n="77"/> si perdeva persino il nome del nostro paese. Ora addio, <persName ref="Lola" type="fictional">Lola</persName>,
<seg type="riferimento_folklorico" subtype="proverbio" source="Demopsicologi.xml#pr_facemu_cuntu"><hi rend="italic">facemu cuntu ca chioppi e scampau,
e la nostra amicizia finiu</hi>.</seg></q></p>
```

Figura 2. Vita_dei_campi.xml

Il proverbio in questione è stato rintracciato anche in una canzone estrapolata dall'opera *Raccolta di proverbi siciliani ridotti in canzoni* del demologo Santo Rapisarda, autore che già Giovanni Battista Bronzini aveva individuato come fonte verghiana certa per i proverbi [1: 294]. Gli studi di Daria Motta sulla lingua di *Vita dei campi* hanno confermato che «La matrice originaria del proverbio è tratta dalla *Raccolta di proverbi siciliani* dell'abate Rapisarda […]» [12: 369].
Come mostra l'immagine (vd. Fig. 3), l'attributo `@xml:id` consente il collegamento tra i due documenti, Vita_dei_campi.xml e Demologi.xml, identificando in maniera univoca il riferimento folklorico in questione
L'esempio è illustrativo, sebbene non esaustivo, del funzionamento delle relazioni tra alcuni dei documenti da analizzare presenti nel *database*. Quando si avrà a disposizione una mole di dati significativa ai fini di un'interpretazione quanto più ampia possibile, i documenti saranno caricati su BaseX, software *open source* che permetterà di interrogare i dati XML e di gestirli a seconda delle diverse esigenze di ricerca.

```
<head>Facemu cuntu ca chioppi, e scampau,
E la nostra amicizia finiu</head>
<cit xml:id="pr_facemu_cuntu" type="riferimento_folklorico" subtype="proverbio">
<bibl>
<title>Raccolta di proverbi siciliani ridotti in canzoni</title>
<author>Santo Rapisarda</author>
<publisher>Giannotta</publisher>
<date>1881</date>
</bibl>
<quote>
<lg>
<l>Amicizia stu cori ti jurau,</l>
<l>E si fu veru amicu lu sacc'iu,</l>
<l>Non cci fu cosa ch'a tia s'ammucciau,</l>
<l>Nè cci fu tra di nui, nè to nè miù:</l>
<l>Ma giacchì senza causa si cangiau</l>
<l>Lu to cori 'ncustanti, o amicu riu,</l>
<l>Facemu cuntu ca chioppi e scampau,</l>
<l>E la nostra amicizia finiu.</l>
</lg>
</quote>
</cit>
```

Figura 3. Demologi.xml

3.1. Un esempio di Georeferenziazione

Per la realizzazione delle edizioni digitali delle opere maggiori di Giovanni Verga, come si è detto, sono state redatte delle indicazioni di codifica comuni che prevedono la marcatura dei luoghi. Una volta ottenuta la `<listPlace>` di ogni opera, si è deciso di georeferenziare su QGIS questi dati, così da integrare allo studio di tipo antropologico un'indagine geografica. Ad ogni punto contrassegnato sulla mappa, corrisponde una lista di campi inseriti nella tabella: "id", "Luogo", "Opera", "Novella" e "Bibl_folkl". Attraverso i campi "Opera" e "Novella" è possibile indicare l'opera e, in maniera più specifica, la novella (o le novelle) in cui Verga menziona un determinato luogo e, attraverso la compilazione del campo "Bibl_folkl" è possibile fornire dei riferimenti bibliografici che rimandino a testi di demologi e folkloristi che nominano quello stesso luogo. L'immagine (vd. Fig. 4) è esemplificativa della struttura del *database* che, nonostante sia stato compilato ancora solo in minima parte, può essere esplorato nella sua progettazione visuale[1].

	id	Luogo	Opera	Novella	Bibl_folkl
1	1	Aci Trezza	Vita dei campi	Fantasticheria	-
2	2	Parigi	Vita dei campi	Fantasticheria	-
3	3	Nizza	Vita dei campi	Fantasticheria	-
4	4	Napoli	Vita dei campi	Fantasticheria	-
5	5	Pantelleria	Vita dei campi	Fantasticheria	-
6	6	Licodia	Vita dei campi	Cavalleria rus...	-
7	7	Sortino	Vita dei campi	Cavalleria rus...	-
8	8	Roma	Vita dei campi	Cavalleria rus...	-
9	9	Canziria	Vita dei campi	Cavalleria rus...	-

Figura 4

L'esempio riguarda le novelle *Fantasticheria* e *Cavalleria rusticana,* i cui luoghi sono stati georeferenziati su QGIS (vd. Fig. 5).

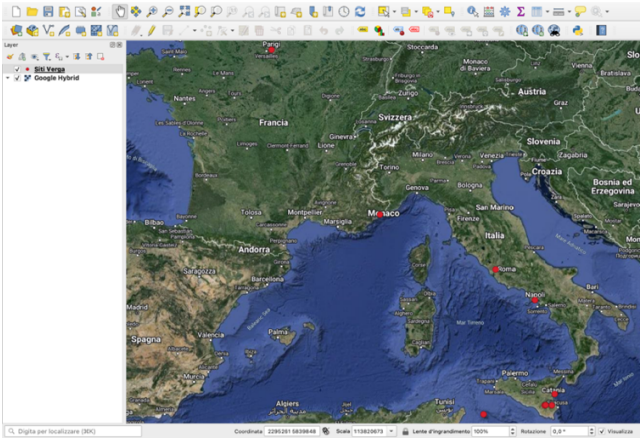


Figura 5

[1] Si noti che, al momento, il campo "Bibl_folkl" è stato tralasciato perché la ricerca si trova in uno stadio iniziale e i riferimenti ai luoghi delle opere demologiche non sono ancora stati ricercati in maniera analitica.

Senza il supporto del *database* la carta geografica con i riferimenti puntuali non darebbe notizie esaustive; avere la possibilità di collegare i riferimenti geografici a dei riferimenti testuali diventa, invece, interessante per l'umanista che voglia trovare corrispondenze tra i luoghi di Verga e quelli dei demologi e studiare il rapporto dell'autore con il folklore.

4. VANTAGGI E IMPATTO ATTESO

Si prevede che un approfondimento del rapporto tra Verga e i primi studi di demopsicologia, anche attraverso l'analisi delle riviste demopsicologiche dell'epoca (Archivio delle tradizioni popolari, Rassegna settimanale, ecc.), possa contribuire a quel dibattito sul folklore che di recente è tornato alla ribalta nel panorama antropologico-letterario. La creazione di un "*database* folklorico" implementabile, contenente i risultati ottenuti dall'analisi geografica e quelli folklorici ricavati dal confronto con le opere demologiche, agevolerà la ricerca e gli studiosi che volessero approcciarvisi. Georeferenziare i luoghi del *corpus* verghiano e porli in relazione con quelli dei testi demologici, fornirà un apporto fondamentale per integrare analisi spaziale e studio dei dati sociali, economici e politici.

L'uso di software *open source* renderà fruibili gli esiti della ricerca a una comunità che non sia solo quella accademica, ma quella di un pubblico più ampio e non necessariamente specializzato e l'utilizzo dell'XML-TEI garantirà l'interoperabilità dei dati.

BIBLIOGRAFIA

[1] Bronzini, Giovanni Battista. «Componente siciliana e popolare in Verga». *Lares* 41, fasc. 3/4 (1975): 275–317. https://www.jstor.org/stable/44630384.

[2] Castellana, Riccardo. *Lo spazio dei Vinti. Una lettura antropologica di Verga*. Roma: Carocci, 2022.

[3] Castellucci, Paola. «Cultura open e cittadinanza scientifica». In *Digital Humanities. Metodi, strumenti, saperi*, a cura di Fabio Ciotti, 214–25. Roma: Carocci, 2023.

[4] Ciotti, Fabio. «La codifica del testo, XML e la TEI». In *Digital Humanities. Metodi, strumenti, saperi*, a cura di Fabio Ciotti. Roma: Carocci, 2023.

[5] Ciula, Arianna, e Cristina Marras. «Modelli, metamodelli e modellizzazione nelle Digital Humanities». In *Digital Humanities. Metodi, strumenti, saperi*, a cura di Fabio Ciotti, 51–65. Roma: Carocci, 2023.

[6] Cristofaro, Salvatore, Christian D'Agata, Antonio Di Silvestro, Giuseppe Palazzolo, Pierpaolo Sichera, e Daria Spampinato. «DEMOTICOn. Per un'edizione semantica dei Malavoglia». In *AIUCD2021 Book of Extended Abstracts*, a cura di Federico Boschetti, Angelo Mario Del Grosso, e Enrica Salvatori, 471–73. Quaderni di Umanistica Digitale, 2021. https://doi.org/10.6092/unibo/amsacta/6712.

[7] Geraci, Mauro. «Quel guardare "da una certa distanza": Verga, il folklore e l'antropologia». In *Verga e il Verismo*, a cura di Giorgio Forni, 217–30. Roma: Carocci, 2022.

[8] Giancristofaro, Lia. *Il segno dei vinti. Antropologia e letteratura in Verga*. Lanciano: Carabba, 2005.

[9] Giuffrida, Milena, Christian D'Agata, Laura Giurdanella, e Pietro Sichera. «Pirandello Nazionale: per un nuovo modello di edizione digitale, collaborativa e integrata». In *AIUCD 2021 - Book of the extended abstracts*, 207–11. Quaderni di Umanistica Digitale. Umanistica Digitale, 2021. https://doi.org/10.6092/unibo/amsacta/6712.

[10] Marotta, Daniele. «Tracciare un'edizione critica digitale. l'esperienza del progetto PhiBor». *Umanistica Digitale* 10 (2021): 89–114. https://doi.org/10.6092/issn.2532-8816/12623.

[11] Marras, Cristina. *Esplorando gli ambienti digitali per la ricerca in filosofia: sostenibilità e modelli di sviluppo*. Video, 19:06. Bologna: Ripresa della conferenza AIUCD2014 – La metodologia della ricerca umanistica nell'ecosistema digitale, 2014. https://youtu.be/brR95gH-UKw.

[12] Motta, Daria. *La lingua fusa. La prosa di Vita dei campi dal parlato popolare allo scritto-narrato*. Acireale: Bonanno, 2011.

[13] Pavia, Davide, Daniela Pasquinelli, e Cristiano Pesaresi. «GIS, geotecnologie e storytelling digitale, tra letteratura e moderna geografia». In *Letteratura e scienze. Atti delle sessioni parallele del XXIII Congresso dell'ADI (Pisa, 12-14 settembre 2019)*, a cura di Alberto Casadei, Francesco Fedi, Annalisa Nacinovich, e Andrea Torre. Pisa: Adi editore, 2021.

[14] Sprugnoli, Rachele, e Timothy Tambassi. «Geografie digitali». In *Digital Humanities. Metodi, strumenti, saperi*, a cura di Fabio Ciotti, 255–66. Roma: Carocci, 2023.

[15] Tomasi, Francesca. «Sistemi informativi e basi di dati». In *Metodologie informatiche e discipline umanistiche*, a cura di Francesca Tomasi, 83–100. Roma: Carocci, 2008.

Progetto di edizione genetica digitale del *Canzoniere* manoscritto di U. Saba (1919-20)

Marina Buzzoni[1], Davide Cucurnia[2], Cristina Fenu[3], Roberto Rosselli Del Turco[4], Giulia Tancredi[5]
[1] Università Ca' Foscari, Venezia, Italia - mbuzzoni@unive.it
[2] Università di Pisa, Italia - davide.i.cucurnia@gmail.com
[3] Biblioteca civica "Attilio Hortis", Trieste, Italia - cristina.fenu@comune.trieste.it
[4] Università di Torino, Italia - roberto.rossellidelturco@unito.it
[5] Università di Siena, Italia - gtancredi94@gmail.com

ABSTRACT
Il «Progetto Saba», avviato su iniziativa della Biblioteca "Attilio Hortis" di Trieste, ha come obiettivo la pubblicazione in forma digitale del manoscritto del *Canzoniere* 1919-1920 di Umberto Saba (1883-1957). Una volta terminata la fase di definizione del modello di codifica XML/TEI più appropriato, il progetto ha visto una decisa accelerazione nel corso del 2023. Sono stati infatti codificati tutti i testi poetici della raccolta e, in parallelo, è cominciato il lavoro di sviluppo per implementare una vista dedicata alla filologia d'autore nel software EVT 3. Questo *paper* intende presentare i risultati preliminari del lavoro fin qui svolto in vista della pubblicazione dell'edizione digitale nella sua forma completa prevista in occasione dell'inaugurazione di Museo LETS - Letteratura Trieste.

PAROLE CHIAVE
Filologia d'autore; Markup XML/TEI; Saba; EVT; Museo LETS.

1. INTRODUZIONE E STATO DELL'ARTE
Grazie a una iniziativa della Biblioteca civica "Attilio Hortis" di Trieste, in collaborazione con l'Università Ca' Foscari di Venezia e altri partner istituzionali[1], e con il contributo di Fondazione Benefica Kathleen Foreman Casali, è in corso dal 2020 un progetto che ha come obiettivo l'allestimento di un'edizione digitale del manoscritto sabiano del *Canzoniere* datato 1919-20 (R.P. Ms I-18) e conservato presso la biblioteca triestina. L'edizione, che sarà pubblicata in modalità *open access* su siti istituzionali dedicati, e accessibile nello spazio espositivo dedicato a Umberto Saba nel Museo LETS – Letteratura Trieste, sarà caratterizzata da funzionalità avanzate di filologia digitale grazie a un'interfaccia dedicata alla filologia d'autore sviluppata all'interno del software EVT[2] (vd. la sezione 3 e [13] per informazioni sulla storia di questo strumento). Il prezioso e inedito manoscritto consiste in un quadernetto di complessive 214 pagine che riporta 186 liriche suddivise in sezioni e sottosezioni.

Ciò che appare particolarmente interessante all'occhio del filologo è che si tratta di una copia predisposta dallo stesso autore per la stampa, come rivelano soprattutto le frequenti notazioni autografe di carattere editoriale. Nel manoscritto risultano evidenti, oltre alle notazioni editoriali, anche cinque fasi correttorie in successione cronologica, contenenti interventi autoriali redatti ciascuno con una penna o matita differente. In particolare, si nota l'uso di: inchiostro blu piombo per la stesura di molte liriche del testo base, una penna a inchiostro blu chiaro, una matita grigia, una matita copiativa, una penna blu oltremare, una matita viola, e infine una penna rossa utilizzata da Saba prevalentemente per fornire indicazioni tipografiche[3].

Alla mano sabiana sono riconducibili anche i numerosi cartigli e pecette adese alle pagine del quaderno. Le varianti d'autore interessano il 67,20% delle liriche e i cartigli il 31,27%.

L'importanza storica di questo testimone del patrimonio culturale italiano, materiale e immateriale, è enorme. Il *Canzoniere* in esso tràdito contiene opere risalenti al periodo 1900-1919 e dunque precede la versione dell'edizione ancora oggi di riferimento, curata da Giordano Castellani, che è basata sulla raccolta pubblicata nel 1921. Si tratta perciò di un documento cruciale sia per la ricostruzione della genesi del *Canzoniere* – di cui tramanda la prima versione in assoluto – sia in quanto riserva ricchissima di varianti, cassature, riscritture e annotazioni che rendono conto dello strenuo processo correttorio cui Saba ha sottoposto la raccolta. L'edizione digitale di R.P. Ms I-18 intende perseguire diversi obiettivi, con lo scopo principale di fare entrare il lettore nel laboratorio poetico sabiano restituendone la complessità dell'incessante processo di

[1] L'elenco completo dei partner comprende, oltre alla Biblioteca "Attilio Hortis" (Trieste) e all'Università Ca' Foscari (Venezia), il Comune di Trieste, l'Università di Torino, il Museo della Letteratura (LETS) di Trieste e il Boston College (Massachusetts, USA).
[2] http://evt.labcd.unipi.it/
[3] *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. P5 Version 4.7.0. Last updated on 16th November 2023. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

riscrittura che lo connota e, dunque, la ricchezza e la stratificazione delle varianti autoriali. In quest'ottica, i curatori hanno a lungo riflettuto su come sia più opportuno modellizzare, codificare e poi visualizzare un'edizione digitale di tipo genetico.

2. DEFINIZIONE DEL MODELLO DI CODIFICA

L'idea di adattare e arricchire gli schemi di codifica TEI per consentire la codifica di edizioni genetiche e di filologia d'autore risale al 2010 circa, e gli strumenti di *markup* disponibili oggi sono il risultato degli sforzi di alcuni studiosi e *editor* TEI particolarmente interessati all'argomento. Il punto di partenza per questa evoluzione degli schemi TEI è il capitolo *Representation of primary sources* delle *Guidelines*[4], e il modulo corrispondente. Su questa base ha cominciato a lavorare un SIG (*Special Interest Group*) TEI che ha prodotto un documento interessante [13], con molte idee valide che sono state poi accettate nel succitato capitolo delle *Guidelines*, ma senza che venisse creato un modulo dedicato alla critica genetica e alla filologia d'autore come auspicato da alcuni dei proponenti.

La mancanza di una "guida" sicura, completa di esempi di marcatura relativi a casi d'uso specifici, ha indotto ogni progetto a creare un proprio modello di codifica sulla base di quanto disponibile nel modulo *core* e in quello di trascrizione delle fonti primarie. Questo ha portato a codifiche molto differenti, e a un certo livello di frustrazione, come si può arguire da questo commento di Britt Barney: "I felt (and feel) less satisfied with this part of the encoding. While it wasn't difficult to describe a step-by-step order of inscription, the nesting of `<seg>` and `<add>` and `<del>` elements in the transcription caused me anxiety." [1]. Restavano inoltre irrisolti gli inconvenienti storici del *markup* XML: "Moreover, the TEI marking, when used for authorial variants, has the drawback of not allowing double marking, tag overlapping […]" [9].

Nel 2020 Eleonora Morante, allieva del Master in Digital Humanities presso l'Università Ca' Foscari Venezia sotto la supervisione di Marina Buzzoni, e Cristina Fenu hanno condotto una prima sperimentazione (non pubblicata) con un *markup* TEI basato sul modulo di trascrizione di fonti primarie e EVT 1: il risultato è stato un'edizione che lascia intravedere le potenzialità del digitale nella resa di annotazioni di filologia d'autore, ma è molto tradizionale nella visualizzazione e gestione delle modifiche autoriali (vd. Fig. 1).

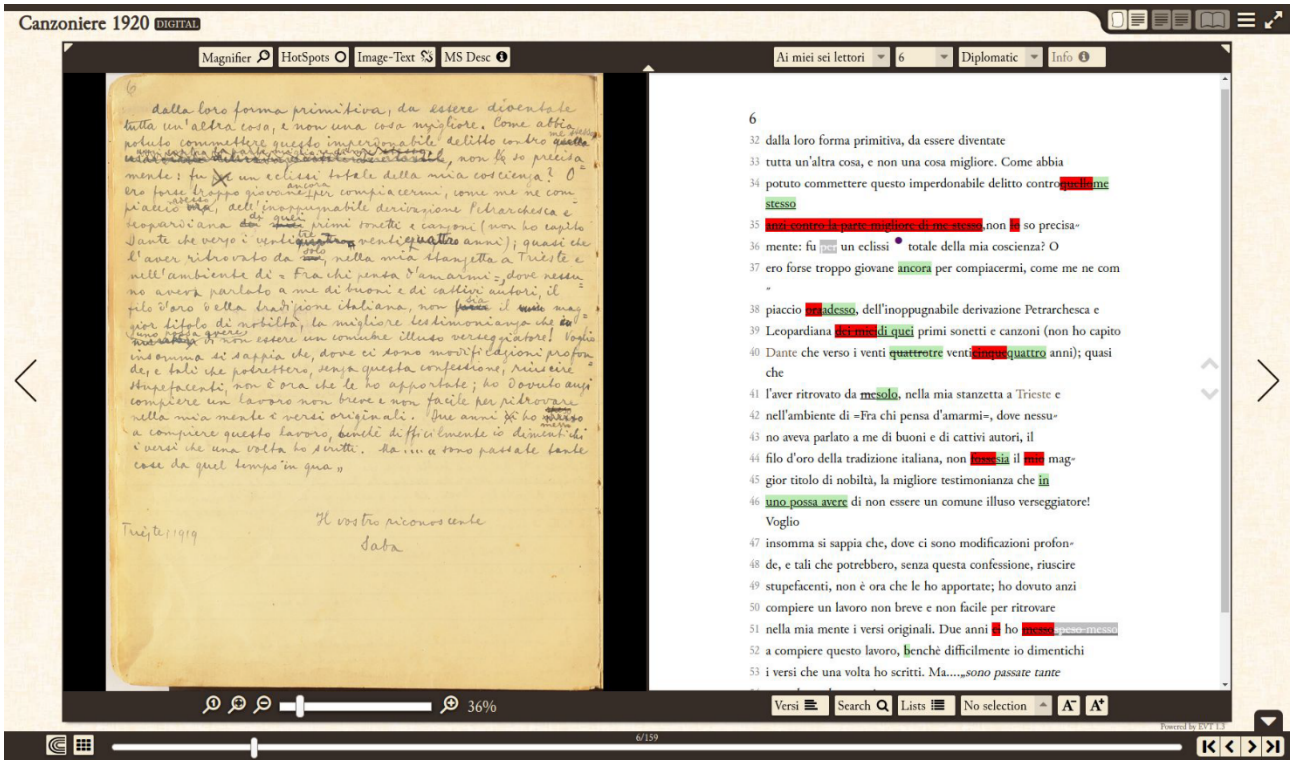


Figura 1. Edizione sperimentale del Canzoniere *usando EVT v. 1.3 (Morante – Fenu 2021)*

Il problema degli schemi TEI nella versione attuale sta soprattutto nel fatto che, mancando un modulo vero e proprio di filologia d'autore, non è stato sviluppato fino in fondo quello che è il punto critico della codifica per le edizioni genetiche, ovvero l'introduzione del fattore tempo nella marcatura. Molto interessante da questo punto di vista il *Proust Prototype*[5] [12], con uno *slider* che permette di riprodurre la sequenza di interventi dell'autore sulle bozze iniziali, offrendo il testo in

[4] http://www.tei-c.org/Guidelines/P5/
[5] http://peterstokes.org/elena/proust_prototype/

box SVG sovrapposti all'area corrispondente nella scansione del manoscritto. Tuttavia è strettamente legato al singolo documento, mentre sarebbe interessante un approccio in grado di seguire l'evoluzione del testo su più manoscritti e/o edizioni a stampa, permettendo di definire un testo critico oltre a un apparato documentale.
L'approccio adottato per l'edizione digitale sabiana qui illustrato è basato su una riflessione generale dello stato dell'arte per quanto riguarda la filologia d'autore digitale e sulla valutazione di progetti che hanno pubblicato edizioni genetiche più o meno sperimentali, in modo da definire un modello di codifica sufficientemente flessibile per edizioni mono- e pluri-testimoniali. Questo ha portato a valutare sia metodi applicabili al documento singolo (modulo di trascrizione di fonti primarie, confronto con gli strumenti adottati per le *documentary editions*), sia gli strumenti offerti dal modulo TEI per edizioni critiche (*Critical Apparatus*, capitolo 12 delle *Guidelines* TEI). Il risultato è una codifica che attinge a più moduli TEI per offrire un apparato basato su fasi e su più livelli testuali, in modo da poter gestire la codifica di casi d'uso diversi, incluse varianti di tipo paradigmatico, e generare un *output* differenziato da un unico documento TEI ("single source publishing"). Saranno pertanto messi a disposizione del lettore

- un apparato trascrizionale, efficace nel registrare le caratteristiche fisiche del manoscritto e delle modifiche apportate nel corso di più campagne correttorie;
- un apparato critico basato sul modulo *Critical Apparatus* che permette di definire un testo critico sia per tradizioni mono-testimoniali, sia nel caso ci siano più testimoni.

I livelli di edizione previsti sono quattro:

- livello **testuale**: le lezioni messe a testo dal filologo (edizione critica), elementi `<app>` e `<lem>` + `<rdg>`
  - in una edizione a testimone unico `<lem>` può indicare l'ultima variante o una emendazione del filologo;
  - in una edizione multi-testimoniale `<lem>` può indicare la variante dell'edizione a stampa o altra selezionata dal filologo, in tal caso è indispensabile l'attributo `@wit` per distinguere testimoni diversi in `<lem>` e `<rdg>`;
- livello **genetico**: le correzioni autoriali successive alla redazione iniziale, codificate tenendo conto della loro sequenza temporale e di altre caratteristiche;
  - le fasi scrittorie sono elencate e descritte in elementi `<change>` contenuti in una `<listChange>` nell'intestazione TEI;
  - nel testo le modifiche sono inserite in elementi `<mod>` collegati alle fasi scrittorie per mezzo dell'attributo `@change`;
  - all'interno di `<mod>` è possibile usare tutti gli elementi di trascrizione: `<add>`, `<del>`, `<metamark>`, ecc.;
  - elementi `<mod>` che rappresentano fenomeni diversi da correzioni immediate sono inseriti in `<lem>` e `<rdg>`, la corretta sequenza temporale è gestita grazie all'attributo `@varSeq` per questi ultimi due elementi, e `@seq` per eventuali `<mod>` annidati;
- livello **intra-testuale**: lezioni che costituiscono varianti alternative, sono codificate per mezzo di `<rdg>` con attributo `ana="#altVariant"`;
  - `@ana` punta a un elemento `<interp>` all'interno di un `<interpGrp>`;
  - particolarmente utili in questo caso attributi come `@place` in `<add>` e altri elementi trascrizionali per indicare la posizione delle varianti alternative;
- livello **meta-testuale**: note e postille autoriali, marcate con l'elemento `<note>` e attributo `@resp` che punta all'autore;
  - attributo `@type` per classificare le note: editoriali, commenti al testo, indicazioni per lo stampatore, citazioni ecc.; ad esempio: `type="metatextual"`, `"autocomment"`, `"quotation"`.

Per una descrizione più articolata, con abbondanza di esempi, si rimanda a [6]. Si noti che la definizione di questo modello ha richiesto molto tempo e, per quanto da considerarsi soddisfacente ai fini della presente pubblicazione, sarà indubbiamente suscettibile di ulteriori modifiche e miglioramenti, anche in base al *feedback* ricevuto dagli addetti ai lavori.

3. EDITION VISUALIZATION TECHNOLOGY V. 3

EVT 3 è un'applicazione web scritta in linguaggio TypeScript e basata sul *framework* Angular per i vantaggi che questi strumenti offrono in termini di supporto allo sviluppo, scalabilità e controllo degli errori. Essendo basato sull'architettura *client-only*, viene eseguito interamente sul computer dell'utente senza la necessità di dialogare e integrare le informazioni con un software di tipo server (ad esempio un database), sia esso locale o remoto. Questo permette un uso flessibile dell'applicativo, riducendo quasi a zero le necessità di manutenzione e i rischi per la sicurezza. Grazie a queste

caratteristiche ha il vantaggio di essere molto facile da installare e configurare, permettendo così di creare rapidamente edizioni digitali sulla base di documenti codificati seguendo lo standard XML/TEI.
L'intera struttura dati dell'applicativo è creata al momento dell'avvio, che in questo caso corrisponde all'apertura della pagina, quando viene eseguito il *parsing* dei documenti XML da visualizzare. Grazie a questo approccio il programma, dopo questa prima fase di avvio, non presenta più tempi di attesa o di caricamento, salvo quelli necessari alla creazione degli elementi sulla pagina e nel DOM tipici degli applicativi con interfaccia web.
Di contro, nel caso di un grande numero di documenti o di documenti di dimensioni considerevoli il programma può presentare un impatto sulla memoria della macchina non trascurabile. Su questo punto sono in corso di sperimentazione varie tecniche di ottimizzazione, quali l'impiego di una struttura dati JSON su cui riversare parte della struttura dati e il caricamento selettivo dei contenuti.
Una prima versione sperimentale di EVT 3 (EVT 3 *alpha version*) è stata pubblicata nel mese di dicembre 2022. Da quel momento lo sviluppo è proseguito su più direzioni diverse, una delle quali è appunto il supporto per la filologia d'autore, con l'obiettivo di arrivare a una versione stabile già nel corso del 2024.

4. SUPPORTO PER EDIZIONI DI FILOLOGIA D'AUTORE IN EVT 3

Una volta definito un modello di codifica sufficientemente stabile per dare inizio a una sperimentazione, il problema da risolvere è stato quello di progettare una interfaccia grafica, da implementare in EVT 3, in grado di gestire i dati della codifica genetica. In breve, gli obiettivi sono i seguenti:

- effettuare un *parsing* del documento XML/TEI per trasformarlo in una struttura dati flessibile e potente grazie al formato JSON, in grado di conservare anche i riferimenti incrociati tra le porzioni di testo;
- trasformare le indicazioni di tipo cronologico (attributi `@varSeq` e `@seq`) in indicazioni efficaci per introdurre il fattore tempo nell'interfaccia utente;
- distinguere tra due modalità di visualizzazione diverse:
    - la prima basata sul livello genetico, quindi inclusiva delle immagini del documento e di una evidenziazione delle modifiche apportate al testo;
    - la seconda basata sul livello critico-testuale, quindi un testo critico con un apparato critico di riferimento per ogni variante;
- sfruttare le caratteristiche di una edizione digitale per permettere all'utente una interazione dinamica con gli apparati e con la gestione dei diversi strati correttori:
    - permettendo all'utente di visualizzare il testo di uno qualsiasi degli strati correttori grazie a un selettore nel riquadro testo (vd. Fig. 2);

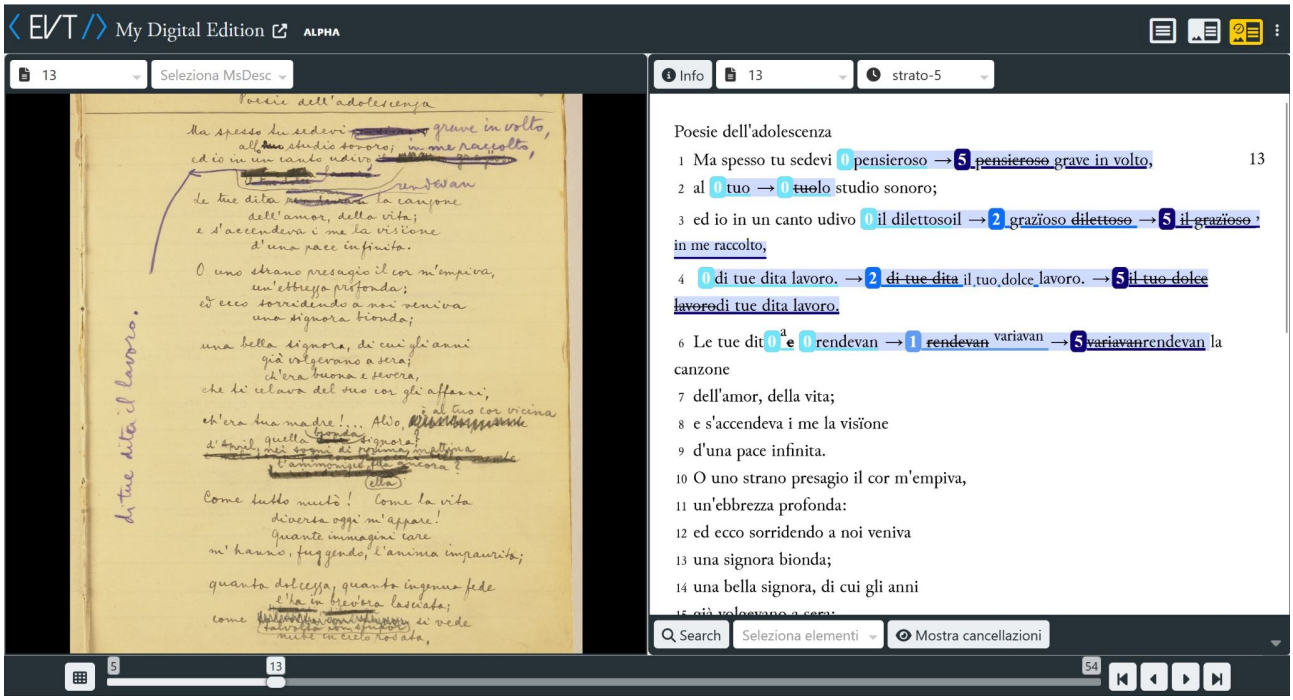


Figura 2. Edizione sperimentale del Canzoniere *usando EVT 3 (Cucurnia 2024)*

    - viceversa, mostrare l'evoluzione del testo per mezzo di una resa visuale di ogni fase come codificata nel documento TEI;
    - aggiungere alla vista critico-testuale la funzionalità di edizione sinottica, in modo da poter comparare il testo delle rispettive fasi;
- visualizzazione migliore delle note autoriali, con collegamento alle corrispondenti aree immagine.
- permettere una certa flessibilità di configurazione della visualizzazione; ad esempio consentendo di mostrare o nascondere su richiesta le cancellazioni o altri elementi informativi.

5. CONCLUSIONI

Il «Progetto Saba» rappresenta un esempio paradigmatico e virtuoso di collaborazione tra studiosi provenienti da ambiti disciplinari diversi, supportati da enti e fondazioni che credono nella capacità della ricerca accademica condotta all'interno del paradigma digitale di sviluppare nuovi modelli per valorizzare il patrimonio culturale su più fronti e in vari modi. Molto rimane ancora da fare: relativamente alla marcatura del testo, l'attività probabilmente più urgente è quella di studiare analiticamente cartigli e pecette, nonché gli strati correttori che li compongono. Sarebbe molto interessante riuscire a recuperare il testo sottoscritto alle pecette adese, ma ciò richiederebbe un delicatissimo intervento sul manoscritto da effettuare in un centro specializzato di recupero e restauro. Si potrebbe sperimentare qualche strumento digitale per il recupero delle scritture cancellate, includendo nel *team* di ricerca specialisti con competenze specifiche in questo ambito di studio. Dal punto di vista applicativo, invece, nei prossimi mesi si lavorerà all'implementazione della parte del progetto maggiormente legata alla fruizione museale dell'edizione che prevede anche l'allestimento di test da somministrare a diverse categorie di utenti organizzati in *focus group* per poter ricevere *feedback* riguardo a fruibilità e accessibilità e intervenire su criticità o ulteriori potenzialità.

BIBLIOGRAFIA

[1] Barney, Brett. «TEI, the Walt Whitman Archive, and the Test of Time». *Journal of the Text Encoding Initiative* 13 (15 maggio 2020). https://doi.org/10.4000/jtei.3249.

[2] Brancato, Dario, Milena Corbellini, Paola Italia, Valentina Pasqual, e Roberta Priore. «VaSto: un'edizione digitale interdisciplinare». *magazén* 1 (2021): 139–69.

[3] Buzzoni, Marina. «Il Progetto Saba: dare voce a un manoscritto inedito del Canzoniere». In *Informatica umanistica, Digital Humanities: verso quale modernità?*, (a cura di) Maristella Gatto, Alessandra Squeo, e Silvia Silvestri, 153–65. Bari: Cacucci Editore, 2024.

[4] Dillen, Wout. «Sequentiality in Genetic Digital Scholarly Editions. Models for Encoding the Dynamics of the Writing Process». In *Digital Humanities 2016. Conference Abstracts. July 11-16, 2016*, (a cura di) Maciej Eder e Jan Rybicki, 174–75. Krakow: Jagiellonian University & Pedagogical University. Alliance of Digital Humanities Organizations (ADHO), 2016. https://doi.org/10.17613/M6GB9B.

[5] Driscoll, Matthew James, e Elena Pierazzo, (a cura di). *Digital Scholarly Editing: Theories and Practices*. Digital Humanities Series. Open Book Publishers, 2016. https://doi.org/10.11647/OBP.0095.

[6] Fenu, Cristina, e Giulia Tancredi. «XML-TEI: Un modello per la filologia d'autore». In *AIUCD 2022 - Culture digitali. Intersezioni: filosofia, arti, media. Proceedings della 11a conferenza nazionale, Lecce, 2022*, (a cura di) Fabio Ciracì, Giulia Miglietta, e Carola Gatto, 218–22. Quaderni di Umanistica Digitale, 2022. https://doi.org/10.6092/unibo/amsacta/6848.

[7] Gabler, Hans W. «Genetic Texts - Genetic Editions - Genetic Criticism or, Towards Discoursing the Genetics of Writing». In *Problems of Editing*, (a cura di) Christa Jansohn, 14:59-78. Beihefte Zu Editio. Tübingen: Max Niemeyer Verlag. De Gruyter, 2012. https://doi.org/10.1515/9783110939958.59.

[8] Italia, Paola. «Filologia d'autore digitale». *Ecdotica* 1 (2019): 203–16. https://doi.org/10.7385/99304.

[9] Italia, Paola, e Giulia Raboni. *What Is Authorial Philology?* Open Book Publishers, 2021. https://doi.org/10.11647/obp.0224.

[10] Muñoz, Trevor, e Raffaele Viglianti. «Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive». *Journal of the Text Encoding Initiative* 8 (28 dicembre 2014). https://doi.org/10.4000/jtei.1270.

[11] Pierazzo, Elena. «Digital Genetic Edition: he Encoding of Time in Manuscript Transcription». In *Text Editing, Print and the Digital World*, (a cura di) Marilyn Deegan e Kathryn Sutherland, Digital Research in the Arts and Humanities:169–86. Surrey; Burlington: Ashgate, 2009.

[12] Pierazzo, Elena, e Julie André. «Autour d'une séquence et des notes du Cahier 46: enjeu du codage dans les brouillons de Proust». In *Colloque « Proust, l'œuvre des manuscrits »*, 2012.

[13] Rehbein, Malte, Lou Burnard, Jannidis Fotis, e Elena Pierazzo. *An Encoding Model for Genetic Editions*, 2010. https://tei-c.org/Vault/TC/tcw19.html.

[14] Rosselli Del Turco, Roberto. «Designing an Advanced Software Tool for Digital Scholarly Editions: The Inception and Development of EVT (Edition Visualization Technology)». *Textual Cultures* 12, fasc. 2 (29 agosto 2019): 91–111. https://doi.org/10.14434/textual.v12i2.27690.

[15] Saba, Umberto. *Il canzoniere: 1921*. (a cura di) Giordano Castellani. Testi e strumenti di filologia italiana. Milano: Fondazione Arnoldo e Alberto Mondadori, 1981.

[16] Vodopivec, Silvia. «Le penne e le matite di Saba. Tracce di volontà autoriale perduta nel Canzoniere (R.P. Ms. 1-18, Biblioteca Civica "A. Hortis" di Trieste).» Università degli Studi di Trieste, 2016.

Rappresentare la Storia Sacra: un'impresa ieri, una sfida oggi. Proposta di edizione scientifica digitale del "Compendium" di Pietro di Poitiers

Franz Fischer[1], Agnese Macchiarelli[2]
[1] Università Ca' Foscari Venezia-VeDPH, Italia - franz.fischer@unive.it
[2] Bergische Universität Wuppertal/Università Ca' Foscari Venezia-VeDPH, Germania/Italia - agnese.macchiarelli@unive.it

ABSTRACT
In questo contributo si presenta e descrive il progetto dedicato all'edizione scientifica digitale del "Compendium historie in genealogia Christi" di Pietro di Poitiers (XII-XIII sec.), un'opera inedita che pone molteplici questioni intorno alla sua rappresentazione. Si introducono inoltre i criteri di edizione delle componenti testuali.

PAROLE CHIAVE
Peter of Poitiers; Compendium historie; Diagrams; Scholasticism; Digital Scholarly Editing.

1. PIETRO DI POITIERS E IL "COMPENDIUM HISTORIE IN GENEALOGIA CHRISTI"
Definito il più fedele allievo del *Magister Sententiarum* [11], Pietro di Poitiers fu attivo come teologo e maestro presso la Scuola Cattedrale di Notre Dame negli anni centrali del XII secolo (m. 1205). La sua opera, vasta e d'impronta scolastica, è nota, ma poco ancora si conosce di un compendio a lui attribuito inerente alla storia narrata nelle Sacre Scritture [13; 22]. Lo scritto in questione, composto entro il 1180, appare come una genealogia che si sviluppa verticalmente e che da Adamo ed Eva giunge al Cristo risorto. Ai nomi, incastonati in medaglioni di norma circolari, si accompagnano glosse testuali di lunghezza variabile che narrano di alcuni episodi intorno ai più importanti personaggi ed eventi che contraddistinguono la discendenza. Concepito come uno strumento per agevolare la comprensione della Storia Sacra (così si legge nel Prologo), il compendio trasmette anche una serie di diagrammi e raffigurazioni che la tradizione restituisce in forme molteplici e spesso rielaborate al fine di venire incontro alle esigenze didattiche degli scolari e di rappresentare efficacemente l'ordine di una storia complessa e universale [17]. Il testimone più antico risale ai primi anni Ottanta del XII secolo (ms. Wien, Österreichische Nationalbibliothek, 363, cc. 1v-6v, 1180-1183[1]), *terminus ante quem* per la datazione dell'opera, e si configura come capostipite di una tradizione la cui portata è ancora in fase di definizione. L'ultimo studio dedicato alla trasmissione del "Compendium" contava circa 270 manoscritti tra rotoli e codici custoditi fin oltreoceano e provenienti in particolare dall'Europa centrale[2] [si vedano anche i precedenti 1, 11, 16: 362–65]. Ricerche più recenti – ora in corso nell'ambito del Progetto "History as A Visual Concept"[3], di cui presentiamo qui le linee principali – hanno accresciuto quel *corpus* di oltre 40 unità, numero già rilevante eppure destinato ad aumentare [3]. L'obiettivo maggiore di questo progetto è infatti quello di fornire un'edizione scientifica digitale di un'opera che di fatto non è mai stata pubblicata criticamente [vd. soprattutto 17 ma cfr. anche 14, 3 e 2 che propone una trascrizione completa, tuttavia poco accurata, del ms. Trivulziano 489] e di cui pure si identificano almeno tre versioni concorrenti che aprono a sfide difficili.
Tuttavia, il "Compendium", pur singolarissimo tentativo di sintesi della Storia Sacra, non è unico nel suo genere. Esistono infatti altre opere, sia letterarie sia storiografiche, parimenti composite e i cui contenuti, diversi per tipologia, forme e soprattutto funzioni, esigono oggi come esigevano allora soluzioni editoriali adeguate. E le riflessioni metodologiche intorno a tali manifestazioni documentarie hanno, per il suddetto motivo, inevitabilmente portato alla creazione di modelli digitali. Del resto, le edizioni digitali consentono di modellizzare grafi informativi molto complessi e permettono agli utenti di diversi settori di confrontarsi con i documenti editati da molteplici prospettive disciplinari e transdisciplinari [15, 13, 14]. Negli ultimi decenni sono apparse numerose risorse, soprattutto nel campo degli studi medievali, e sin dalle prime pubblicazioni, divenute poi di riferimento come quelle su CD-ROM del "Wife of Bath's Prologue" di Chaucher [6] e del "Beowulf" [10], i metodi e i formati delle rappresentazioni digitali dei testi sono stati costantemente sviluppati e ampliati, e con essi anche la loro portata. La trasparenza delle scelte editoriali, l'utilità delle edizioni insieme con le funzionalità

[1] https://digital.onb.ac.at/RepViewer/viewer.faces?doc=DTL_4587063&order=1&view=SINGLE.
[2] Piggin, Jean-Baptiste. *Peter's Stemma*. https://www.piggin.net/stemmahist/petercatalog.htm
[3] DFG/FWF-funded Research Project „Geschichte als visuelles Konzept: Peter von Poitiers' Compendium historiae" / "History as a Visual Concept: Peter of Poitiers' Compendium Historiae", 2023-2025, https://fit.uni-tuebingen.de/Project/Details?id=10031.

sono state quindi accresciute, ad esempio, fornendo facsimili, trascrizioni diplomatiche complete e strumenti critici, quali software di collazione o di visualizzazione delle strutture [8, 7, 9].

2. COME RAPPRESENTARE DIGITALMENTE IL "COMPENDIUM"

Mentre la maggior parte delle edizioni e degli strumenti esistenti prendono in considerazione la sola componente testuale, alcune iniziative editoriali, come la "Bayeux Tapestry Digital Edition" o l'"Emblem Project Utrecht" o il recente "Welscher Gast Digital", si concentrano anche sulla dimensione visuale e materiale dei manufatti medievali e, in particolare, sulle relazioni testo-immagine-oggetto. Questi progetti si basano però su un rapporto convenzionale testo-immagine, in cui l'immagine è considerata come un supplemento illustrativo alla narrazione oppure, al contrario, il testo è visto come nota di commento all'immagine. Per le opere con una forma diagrammatica intrinseca come il "Compendium", le suddette componenti non possono essere separate, né gerarchizzate. Pertanto, nel rappresentare l'opera di Pietro di Poitiers, occorre considerare e valorizzare ogni singolo aspetto: strutturale, testuale, visuale/iconografico, diagrammatico, semantico.
Nel complesso, tali componenti insieme con le unità che le costituiscono (linee/connessioni; nomi/nodi e glosse/blocchi testuali; figure; diagrammi; rubriche e didascalie) rappresentano le entità di una sovrastruttura i cui livelli sono determinati a partire da un'ontologia definita su base relazionale. Attraverso la sovrastruttura, è quindi sviluppata la risorsa al fine di: 1. visualizzare la struttura grafica della genealogia secondo il formato SVG; 2. stabilire un'edizione di riferimento codificata in XML/TEI[4] [4]; 3. visualizzare l'intera opera tramite un'interfaccia navigabile, online e fruibile da un pubblico di esperti e non solo; 4. annotare le immagini/facsimili dei manoscritti nel rispetto del manifesto IIIF[5] alle quali associare, in primo luogo, le trascrizioni delle componenti testuali. Al prodotto finale si intende poi collegare un database creato per censire e descrivere i testimoni; classificare le varie versioni dell'opera mediante l'individuazione di glosse testuali distintive; offrire un'interpretazione storico-artistica delle figure, dei diagrammi e delle loro varianti.

3. VERSO L'EDIZIONE: LE COMPONENTI TESTUALI

All'interno della tradizione del "Compendium" si distinguono con certezza almeno tre versioni. La prima, attestata dalla maggioranza, è la più breve e parrebbe corrispondere a quella originale. È composta da circa 450 nodi, 160 blocchi testuali, 5 diagrammi e un numero di immagini variabile a seconda della fattura e della destinazione d'uso del manoscritto (cfr. il ms. di Vienna [vd. Figg. 1a e 1b]; oppure il ms. Klosterneuburg, Augustiner-Chorherrenstift, 696[6] [vd Fig. 2]).

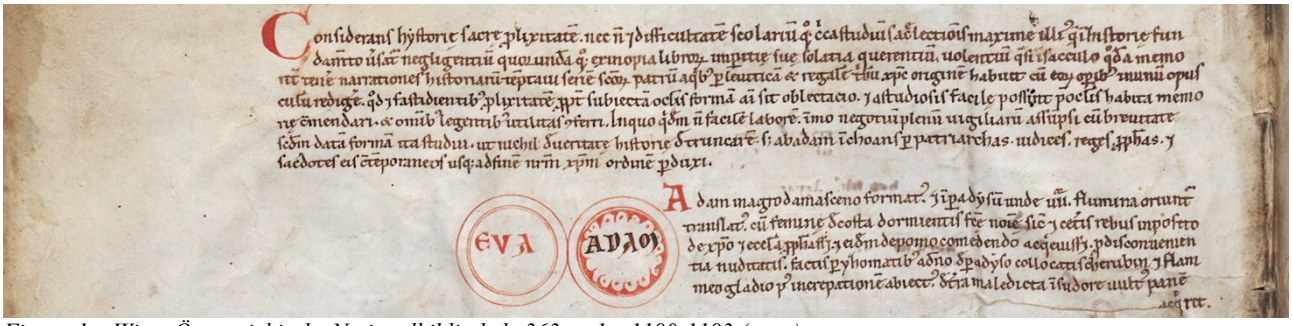


Figura 1a. Wien, Österreichische Nationalbibliothek, 363, c. 1v, 1180-1183 (part.)

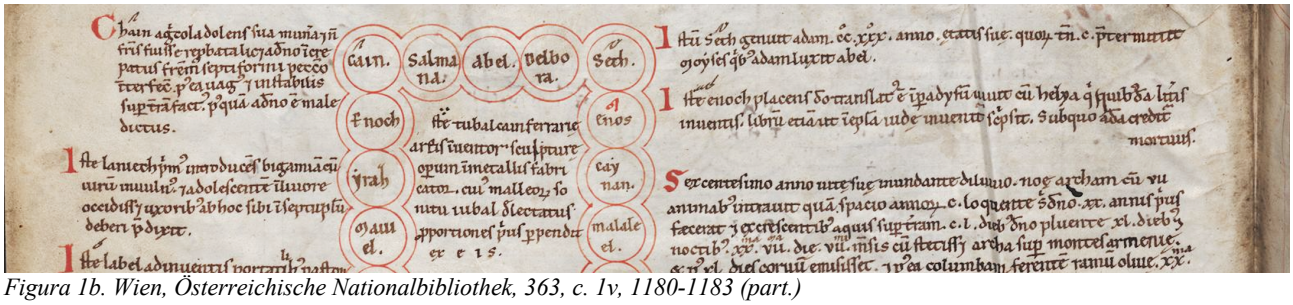


Figura 1b. Wien, Österreichische Nationalbibliothek, 363, c. 1v, 1180-1183 (part.)

[4] https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html
[5] https://iiif.io/
[6] https://manuscripta.at/diglit/AT5000-696/0009?sid=f576a4ea7c1403f776916bf735f0e94a.

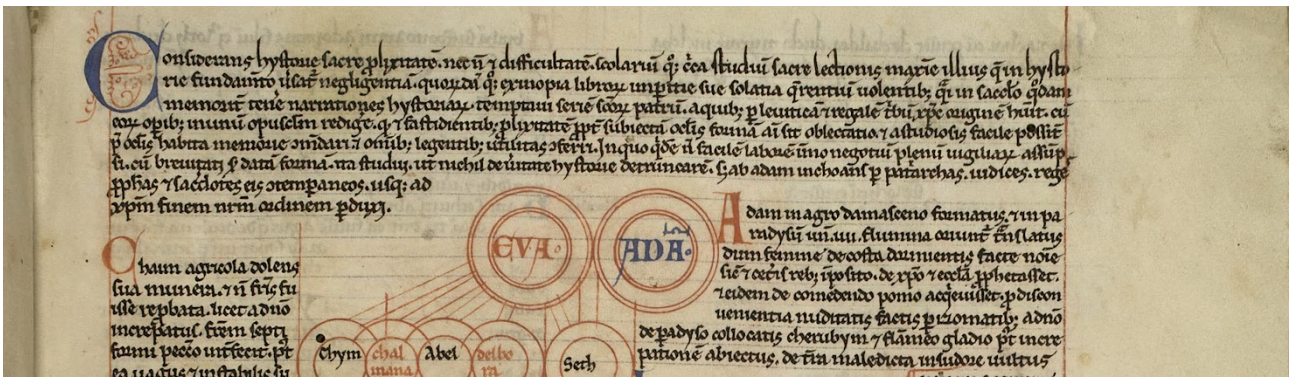


Figura 2. Klosterneuburg, Augustiner-Chorherrenstift, 696, c. 2r, sec. XIII (part.)

La seconda versione, definita 'interpolata', si allontana dalla prima per l'aggiunta di una quindicina di blocchi testuali e l'omissione di pochi altri. È trasmessa, già in epoca antica, per esempio dal rotolo di Cambridge (US) Harvard College Library, Typ 216[7], datato alla prima metà del 1200 (vd. Fig. 3) o da quello di Cleveland ascrivibile al sec. XIII.

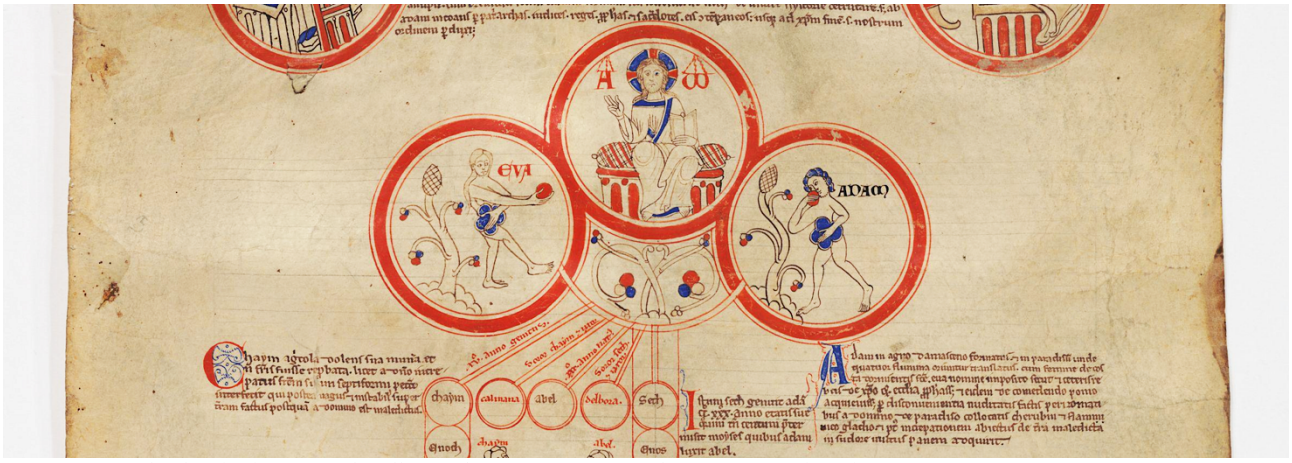


Figura 3. Cambridge (US), Harvard College Library, Typ 216, rotolo, 1200-1250 (part.)

La versione cosiddetta lunga è riconoscibile a colpo d'occhio per estensione. Tuttavia, a differenza delle altre pressoché stabili, è la più mutevole. Oltre a diffondere porzioni testuali nuove, dove sono menzionate esplicitamente le fonti dell'opera, amplia non solo le glosse originali di episodi in precedenza omessi, ma accresce anche la genealogia canonica di linee regali locali che riflettono il luogo di provenienza dell'esemplare e completa la lista dei pontefici noti fino a quel determinato momento (cfr. fra tutti il ms. Paris, Bibliothèque de l'Arsenal, 1234[8] [vd. Fig. 4]).

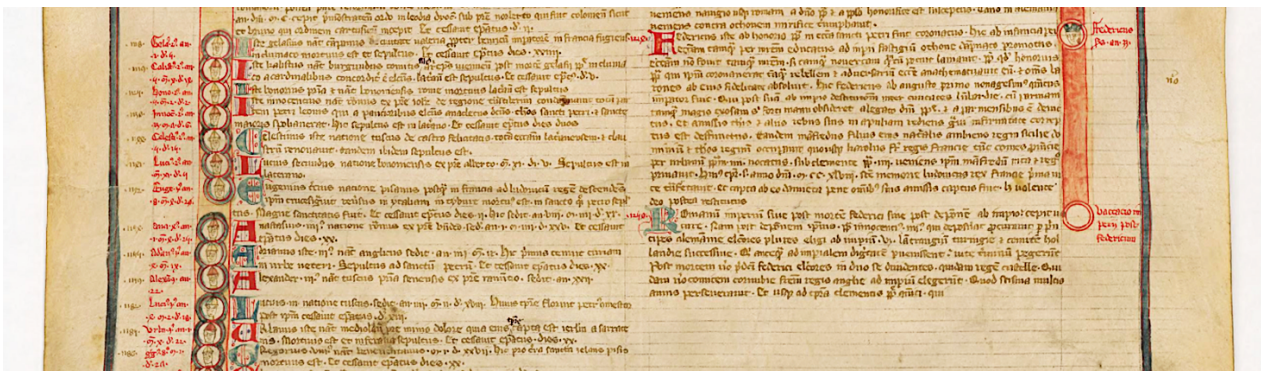


Figura 4. Paris, Bibliothèque de l'Arsenal, 1234, rotolo, sec. XIV (part.)

[7] https://iiif.lib.harvard.edu/manifests/view/drs:12438364$1i
[8] https://gallica.bnf.fr/ark:/12148/btv1b525067595/f1.item.zoom

Il quadro dunque, pur chiaro, è assai volubile e il tasso di variazione fra le versioni e all'interno delle stesse è alto. Pertanto, ai fini dell'edizione del "Compendium", si è scelto di pubblicare integralmente solo la versione breve, la cui struttura, come detto, è nel complesso regolare.

Sulla base dei testimoni più antichi (*ante* 1250), fra i quali spiccano il suddetto codice viennese (W) e il ms. Cambridge, Corpus Christi College, 29 (sec. XIII, primo quarto)[9], verrà stabilito il testo di riferimento, codificato secondo il modello XML/TEI. Nella prima fase, ormai conclusa, sono state prese in considerazione anzitutto le componenti testuali dell'opera e cioè nomi (*nodes*), glosse (*textblocks*) e, laddove non vi fosse una variazione importante, didascalie (*labels*) ed elementi testuali racchiusi nei diagrammi (*diagrams*), identificati tramite un PID. Di queste è stata approntata una trascrizione semidiplomatica basata sul ms. di Cambridge, più vicino al disegno originario rispetto a W. La seconda fase è ora in corso e ha come obiettivo quello di stabilire il primo testo di riferimento normalizzato secondo specifiche regole e corretto a partire dai risultati della collazione condotta sulla restante parte dei testimoni selezionati. Tale forma testuale costituirà il punto di partenza per una visualizzazione immediata del testo all'interno del grafico navigabile. Sempre in questa fase, tenuto conto della *varia lectio* registrata, si produrranno – in modo automatico – anche le trascrizioni normalizzate di tutti i testimoni interessati al fine di associare tali versioni di lettura all'immagine facsimilare del manoscritto corrispondente[10].

In un terzo momento, il testo di riferimento verrà corredato di un apparato critico codificato secondo il "Parallel segmentation method"[11], di un apparato delle fonti bibliche e secondarie e infine di note esplicative. Le altre componenti sono oggetto di studi paralleli i risultati dei quali confluiranno nei rispettivi luoghi e livelli adibiti alla loro visualizzazione all'interno dell'interfaccia navigabile. È programmata anche una traduzione in lingua inglese che insieme con un'introduzione generale completerà l'edizione della versione breve.

A un secondo livello, saranno integrate anche le trascrizioni semidiplomatiche delle glosse aggiuntive peculiari della versione interpolata. Per ciò che concerne la versione lunga, assai rielaborata nel corso della tradizione e che necessita di studi specifici anche sui singoli testimoni, si intende invece rimandare al database, dove sarà possibile trovare una descrizione analitica dei manoscritti che la tramandano e talune informazioni utili al suo riconoscimento.

4. CONCLUSIONI

Pensare di rappresentare in maniera tradizionale il "Compendium" di Pietro di Poitiers è un'impresa che non darebbe frutti. Come un'impresa fu per lo stesso Pietro pensare di riassumere lungo cinque metri o poco più l'intera genealogia di Cristo con i suoi fatti, le sue leggende, le sue persone. Ma in quel caso, a differenza di quanto potrebbe accadere se si volesse stampare oggi il "Compendium", di frutti ce ne furono... e pure buoni. Esso conobbe infatti una straordinaria fortuna, divenendo sin dai primi anni della sua diffusione non solo 'testo' didattico ma anche 'testo' di corredo di opere di chiara fama, come la "Historia scholastica" di Pietro Comestore a cui nella tradizione spesso si accompagna. Eppure, la complessità strutturale e la ricchezza di contenuti dell'opera di Pietro di Poitiers hanno fatto sì che non fosse mai pubblicata criticamente, assunte, forse, le concrete difficoltà di giungere a una versione univoca il più possibile vicina all'ultima volontà dell'autore (primo e unico tentativo fu quello di Zwingli nel 1592 [19]). La sfida per l'editore moderno è dunque quella di rappresentare tutti gli aspetti e le componenti di un'opera poliedrica senza denaturalizzarla: operando, cioè, scelte ragionate che non privino l'opera della(e) sua(e) vera(e) essenza(e). Ricorrere al digitale è pertanto l'unico modo attraverso cui pervenire, dopo lunga modellizzazione, a un punto di riferimento scientificamente determinato che però rifletta la stratificazione e la dinamicità di un'opera nata, in apparente contrasto, per fissare.

BIBLIOGRAFIA

[1] Alidori, Laura. «Il Plut.20.56 della Laurenziana. Appunti sull'iconografia dei manoscritti della "Genealogia" di Petrus Pictaviensis». *Rivista di Storia della Miniatura* 6–7 (2002 2001): 157–70.

[2] Baroni, Maria Franca. *Un prezioso rotolo storico religioso del sec. XIII*. Milano, Varese: Istituto editoriale cisalpino, 1969.

[9] https://parker.stanford.edu/parker/catalog/xj710dc7305.

[10] La "TEI - Text Encoding Initiative, ossia lo standard per le edizioni scientifiche digitali, fornisce solo alcune raccomandazioni per la codifica delle immagini, il collegamento testo-immagine e la rappresentazione di grafi e network (cfr. i capp. "11. Representation of Primary Sources", in particolare i parr. "11.1. Digital Facsimiles" e "11.2. Combining Transcription with Facsimile"; si veda anche il cap. "19. Graphs, Networks, and Trees"). Pertanto, il progetto sul "Compendium" non costituirà solo un banco di prova per la loro applicazione e il loro ulteriore miglioramento, ma darà anche un importante contributo alla ricerca e alla discussione in corso sulla rappresentazione digitale di testi, immagini e, più in generale, conoscenza.

[11] Cfr. il cap. "12.2.3. The Parallel Segmentation Method" delle TEI *Guidelines*, https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS.

[3] Bleier, Roman, Laura Cleaver, Elisa Cugliana, Eleanor Goerss, Franz Fischer, Sina Krottmaier, Agnese Macchiarelli, Patrick Sahle, Maria Streicher, e Andrea Worm. «History as a Visual Concept: Peter of Poitiers' "Compendium historiae». *Manuscript Studies* 9, fasc. 2 (Fall 2024, in press).

[4] Bleier, Roman, Franz Fischer, Tessa Gengnagel, Patrick Sahle, e Andrea Worm. «Session Fri 3b: TEI and models of text III. Paper 1: Text - Graph - Image: Towards a Digital Edition of Peter of Poitiers' Compendium historiae». In *What is Text, Really? TEI and Beyond, September 16–20, University of Graz, Austria, Book of Abstracts*. Graz, 2019.

[5] Bollati, Milvia. «Simboli e diagrammi nel "Compendium historiae in genealogia Christi" di Pietro di Poitiers: la Menorah». In *Ordinare il mondo. Diagrammi e simboli nelle pergamene di Vercelli.*, (a cura di) Timoty Leonardi e Marco Rainini, 211–32. Milano: Vita e Pensiero, 2018.

[6] Chaucer, Geoffrey. *The Wife of Bath's Prologue on CD-ROM. The Canterbury Tales Project*. (a cura di) Peter Robinson. Cambridge: Cambridge University Press, 1996.

[7] Fischer, Franz. «Digital Classical Philology and the Critical Apparatus». In *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, (a cura di) Monica Berti, 203–20. Berlin/Boston: De Gruyter, 2019.

[8] Fischer, Franz. «Digital Corpora and Scholarly Editions of Latin Texts: Features and Requirements of Textual Criticism». *Speculum* 92, fasc. S1 (2 ottobre 2017): 265–87.

[9] Fischer, Franz. «Representing the Critical Text». In *Handbook of Stemmatology. History, Methodology, Digital Approaches*, (a cura di) Philipp Roelli e Aidan Conti, 405–27. Berlin, Boston: De Gruyter, 2020.

[10] Kiernan, Kevin. *Beowulf and the Beowulf Manuscript*. Ann Arbor: University of Michigan Press, 1996.

[11] Moore, Philipp Samuel. *The Works of Peter of Poitiers. Master in Theology and Chancellor of Paris (1193–1205)*. Notre Dame: Publications in Medieval Studies, 1936.

[12] Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods*. Digital Research in the Arts and Humanities. Farnham Surrey: Ashgate, 2015.

[13] Pierazzo, Elena, e Peter Stokes. «Putting the Text back into Context: A Codicological Approach to Manuscript Transcription». In *Codicology and Palaeography in the Digital Age 2*, (a cura di) Franz Fischer, Christiane Fritze, e Georg Vogeler, 397–430. Norderstedt: BoD, 2010.

[14] Rainini, Marco. «I rotoli del "Compendium historie in genealogia Christi" di Pietro di Poitiers: origini e primo sviluppo dal testimone di Milano, Biblioteca Trivulziana, ms. 489». In *Imago librorum: mille anni di forme del libro in Europa. Atti del convegno di Rovereto-Trento, 24–26 maggio 2017*, 41–77. Firenze: Leo S. Olschki, 2021.

[15] Sahle, Patrick. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung. [Finale Print-Fassung]*. Vol. 9. Norderstedt: BoD, 2013.

[16] Stegmüller, Friedrich. *Ed. Repertorium Biblicum Medii Aevi*. Vol. 4 Commentaria: auctores N–Q. Madrid: Consejo Superior de Investigaciones Científicas. Instituto Francisco Súarez, 1954.

[17] Worm, Andrea. *Geschichte und Weltordnung: Graphische Modelle von Zeit und Raum in Universalchroniken vor 1500*. Berlin: Deutscher Verlag für Kunstwissenschaft, 2021.

[18] Worm, Andrea. «Visualising the Order of History: Hugh of Saint Victors' "Chronicon" and Peter of Poiters' "Compendium Historiae». In *Romanesque and the Past: Retrospection in the Art and Architecture of Romanesque Europe*, (a cura di) Richard Plant e John McNeill, 243–63. Leeds: Maney, 2013.

[19] Zwingli, Ulrich. *Petri Pictaviensis Galli genealogia et chronologia sanctorum patrum*. Basileae: Per Leonhardum Ostenium, 1592.

Towards an integrated digital edition of the *Leges Langobardorum*

Marina Buzzoni[1], Roberto Rosselli Del Turco[2]
[1] Università Ca' Foscari, Venezia, Italia - mbuzzoni@unive.it
[2] Università di Torino, Italia - roberto.rossellidelturco@unito.it

ABSTRACT

The *Leges Langobardorum* project, born within the ALIM project (PRIN-2012) and now founded in the framework of PRIN PNRR 2022, builds on the current state-of-the-art to provide an integrated edition of this law texts corpus, handling both separate diplomatic transcriptions and "virtual witnesses" automatically generated thanks to the TEI markup for critical editions. It also aims to provide a seamless navigation between critical text, diplomatic transcriptions and digital facsimiles. Furthermore, the visualisation software EVT will be equipped with accessibility settings, i.e. features that let people with disabilities customise a device for their own needs, which can represent a true life-changer for people with special needs.

KEYWORDS

Leges Langobardorum; Digital Philology; Textual Criticism; XML/TEI; EVT.

1. INTRODUCTION AND STATE-OF-THE-ART

The original idea of the *Leges Langobardorum* project, now founded in the framework of PRIN PNRR 2022, was born in 2015 as a collaboration between the Research Units (PRIN-2012) of the ALIM - "Archivio della Latinità Italiana del Medioevo" project[1]. Initially limited to the text of the *Edictus Rothari*, which is also the testbed of the current experimental editions, it was later extended to cover the whole body of Lombard laws as preserved in several manuscripts and manuscript fragments. The Lombard culture, as well as its relationship with different forms of writing, particularly important in the moment they decided to write down their *Leges*, is a real building block of Italian culture within the wider European setting. The need for a new, digital edition of this corpus of laws, crucial for the studies of the Lombard language and for a better appreciation of their culture, arises from two different necessities: the inadequacy of the existing critical editions and the impossibility of publishing an edition including facsimiles of the manuscripts, their diplomatic rendering, as well as the critical edition except in digital form. The codices that preserve the *Leges Langobardorum* amount to a dozen (counting the Codex Heroldinus which has survived in a printed edition of 1557) and span a wide chronological arc between the second half of the seventh century and the eleventh century. To these must be added a series of fragments (14) that preserve parts of the law texts, thus increasing the total number of witnesses, currently estimated at 26. In order to reconstruct the documentary history of the *Leges*, it is also necessary to take into account the three Lombard-Latin glossaries of the 10th/11th-13th- c. (Cavense, in ms 9; Matritense, in ms 8; Vaticano, probably from Salerno), which include the interpretation of terms also taken from the juridical corpus [13].

The reference edition for the entire corpus of "Lombard Laws" is still the one prepared by Bluhme in 1868 for the *Monumenta Germaniae Historica*. In 1947 Beyerle published *Die Gesetze der Langobarden*, reprinted in 1962. More recently, Azzara [2] offers an important translation of the text of the Laws, to which he adds an edition of the original that builds on Beyerle, but departs from it in some points.

A new edition of the *Leges* is definitely needed for a number of reasons:

(1) the editions currently available to scholars are all based on ms. 1 (Codex Sangallensis 730), the oldest one, which does not always transmit Lombard terms in accurate form;

(2) a complete *recensio* of the extant witnesses has never been carried out;

(3) it would be enlightening to take the fragments into account more than has been done so far, including the most recently discovered ones;

(4) the editions produced to date (cited above) have been judged not to be satisfactory by experts in Lombard language and culture;

[1] http://it.alim.unisi.it/

(5) the new edition would be published in a digital form that is potentially more inclusive since it would reach a larger and more differentiated audience, and used in fields not necessarily strictly related to academia (e.g. tourism, museums and, more generally, the GLAM sector).

As for the drawbacks of the editions currently available, Molinari [16: 234] considers the text restored by Bluhme "readable, but a-historical in terms of both language and content." Beyerle [3] builds upon Bluhme's text, but does not take into account the fragments discovered by Dold in the 1930s [7: 1-52]. Azzara [7: 1-52] uses Beyerle as a base, but departs from it at specific points because of unspecified "textual criticisms that were deemed necessary." The author seems well aware of the substantial fragility of his editorial method when he warns the reader that: "[t]he present edition represents only a start, a first step, toward a future critical edition" [2: lx].

We also strongly believe that a well-designed scholarly digital edition is the best method to publish a new, more up-to-date and philologically correct edition of the *Leges*. An edition that starts from the diplomatic-interpretive rendering of the manuscripts, also providing digital images, would be able to represent the historical dimension of each witness much more accurately than a printed edition. Furthermore, a digital edition can facilitate the study of Lombard terms, and their Latin glosses when present, not in isolation, but in their context of transmission. Moreover, the iconographic programme in the tradition of illuminated manuscript witnesses of the "Lombard Laws" bears a crucial historical value in itself. The 'southern' codices (Cavensis and Matritensis), for example, contain a number of images depicting the legislating kings and prince-legislators (see [8, 12, 22]). In fact, only a digital environment can efficiently highlight the peculiarities of the text-image relationship in these codices.

2. GENERAL PROJECT ROADMAP

The overall project is divided into the following phases:

(1) The first phase, which will be conducted in parallel with 2 and 3, will see a reassessment of the textual tradition of the *Leges* to ascertain which witnesses and fragments are needed in order to complete the *recensio* step, necessary for the preparation of the critical edition. The identified documents will be likewise digitized, so that they can be included in the *constitutio textus* phase.

(2) Phase 2 aims at providing a diplomatic-interpretive edition of the "Piedmontese" witnesses to the *Leges*, namely: ms 2 Codex Vercellensis (Vercelli, Biblioteca Capitolare CLXXXVIII, see Fig. 1) and ms 3 Codex Eporedianus (Ivrea, Biblioteca Capitolare, XXXIV).

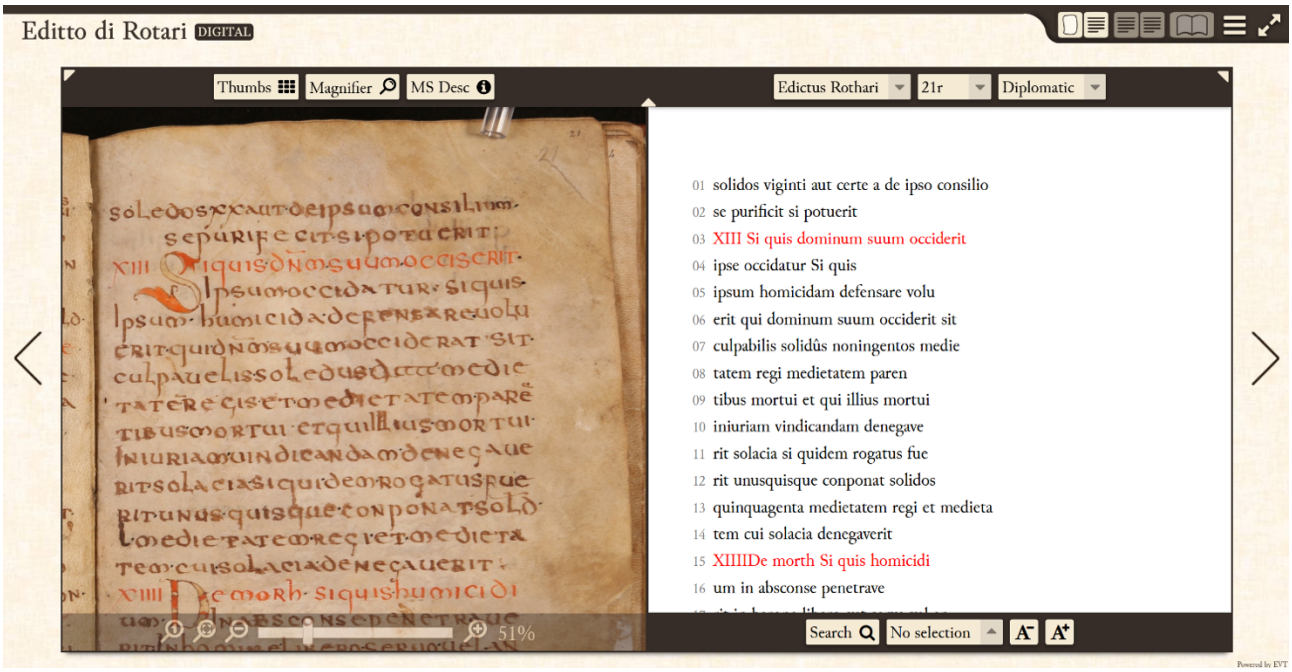


Figure 1. The first experimental edition of the Codex Vercellensis *CLXXXVIII using EVT 1*

(3) In the third phase, the same process will be repeated for witnesses identified in phase 1, including the fundamental ms 1 Codex Sangallensis (St. Gallen, Stiftsbibliothek, Cod. Sang. 730) and ms 5 Codex Vaticanus (Biblioteca Apostolica Vaticana, Vat. Lat. 5359) whose scanned images are available online.

(4) The critically defined texts of all selected witnesses will then flow into the ALIM digital library, while full digital editions (facsimile + transcriptions) will be published using a dedicated software: EVT (Edition Visualization Technology).
(5) The fifth, more laborious phase involves the preparation of the critical edition of the text of the *Edict* based on the comprehensive collation and review of the witnesses which have been selected for that purpose.
(6) In the final phase, the EVT software will be developed to include an integrated edition feature: an innovative functionality which will allow it to handle digital facsimiles, diplomatic transcriptions and a critical edition enabling the user to seamlessly navigate from one to another, e.g., linking variant readings to the corresponding form in the context of the reference witness, and from there move to the digital facsimile to examine the material evidence of the same form.
Phase 2 has already been developed and completed as part of the ALIM project, while phase 1 and 3 are underway. An important step will be a thorough check of the transcriptions both on the linguistic (mediaeval Latin) and the technical (XML encoding) level prior to their publication.

3. INNOVATIVE ASPECTS AND METHODOLOGIES

The most important innovative aspects of our project are:

- The idea of providing an **integrated edition**, handling both separate diplomatic transcriptions and "virtual witnesses" automatically generated thanks to the TEI markup for critical editions (see Fig. 2);
- The handling of named identities within a TEI-compliant glossary (see Fig. 3);
- The full digitization of the most important witnesses, including the fragments, which are of high philological value;
- A new EVT 3 feature, themes, will be used to implement support for people with disabilities such as colour-blindness.

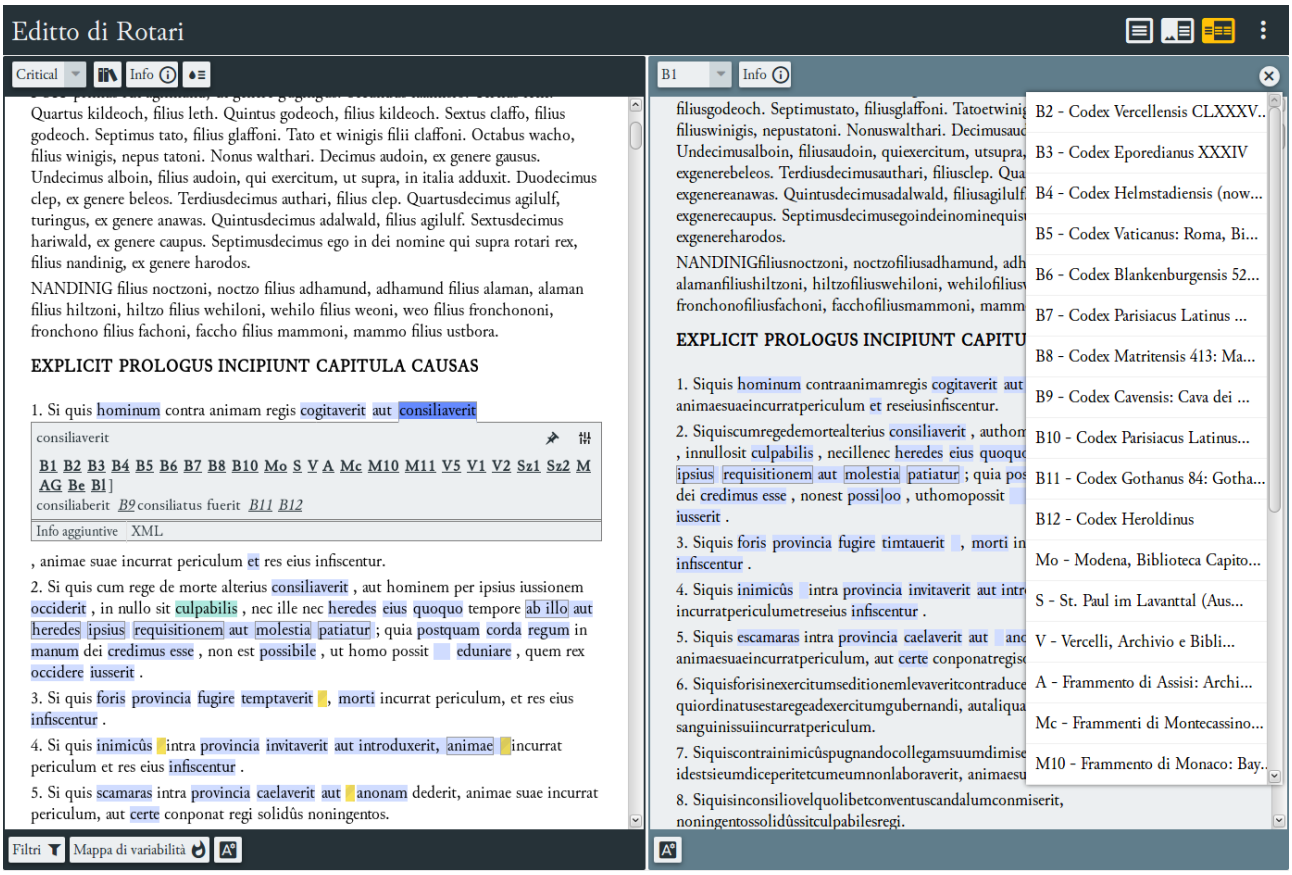


Figure 2. The critical edition of the Edictus Rothari *in EVT 2 with a list of automatically generated witnesses on the left*

For codices included in the project, but which are not already available because of previous acquisitions or thanks to publication by digital libraries (with a preference for servers relying on the IIIF framework), a digitization is planned to be carried out either at the DH Center 'Digital Scholarship for the Humanities' of the Turin Unit (DISH[2]) or on site thanks to portable scanners again provided by DISH. For texts lacking a transcription, a new one will be carried out by means of the

[2] https://www.dish.unito.it/

TEI[3] schemas and *Guidelines* (TEI Consortium 2023[4]) based on the XML markup language, an established standard in the production of digital editions, in order to prepare a digital facsimile accompanied by the corresponding transcription.

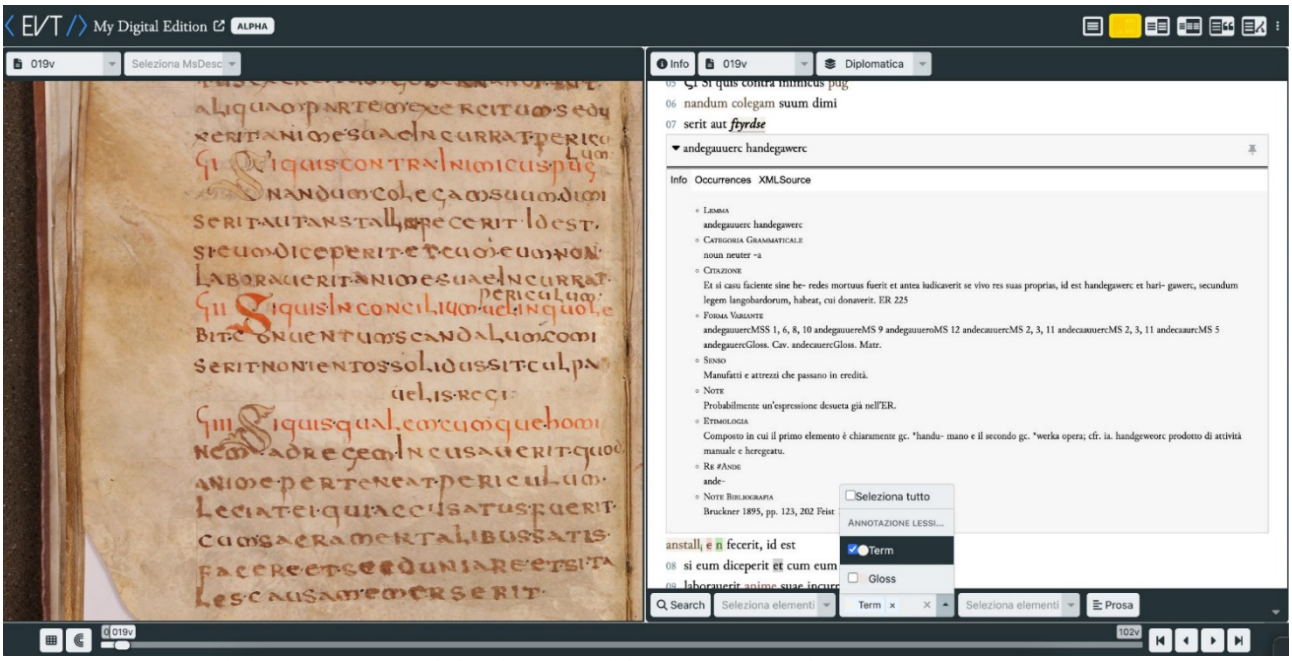


Figure 3. Experimental glossary support added to the EVT 3 alpha version (December 2022)

The newly digitised materials, and the XML/TEI documents containing the texts of the respective editions, will also be accompanied by metadata so that the components of the global edition can be documented in view of their subsequent dissemination.

The diplomatic editions will be published by means of the EVT software [20], whose graphical interface and display system allow for a more immediate interaction with the images of the codices, the texts of the digital edition, and the supporting and commentary materials that will gradually be included within the various editions. The solid performance of such mature software with TEI-based editions has already been extensively proved by its use in projects such as the Digital Vercelli Book[5] [21], the Codice Pelavicino Digitale[6] [23], the Petrus de Ebulo critical edition[7] and the DI edition of Marco Polo's *Devisement dou Monde*[8].

The use of an open source and flexible system such as EVT, moreover, will enable the development of a set of specific features necessary to achieve the goals of the project: in particular, adding support for critical editions accompanied by the images of a specific witness and the development of an integrated edition system within EVT. The latter is especially important because, at the time of writing, diplomatic editions alongside digital facsimiles and critical editions reflect two distinct methodological approaches that are functional to their respective goals, but in fact prevent a "dialogue" even within the same textual tradition. They are in fact the result on the one hand of the new philology [18] and digital documentary editions [19] approach, which is predominant in the Anglo-Saxon world, and on the other hand of the traditional stemmatic method, which can boast an abundant tradition in continental Europe and, in particular, in Italy. As a result, each scholar follows the method he or she considers most appropriate, but the final products, however good their quality, are "islands" that do not communicate with each other. An integrated edition, on the other hand, constitutes an important methodological innovation precisely because it would allow both the diplomatic editions of the most relevant witnesses and the critical edition that is produced on the basis of their collation to be brought under a single navigation and research environment, one that would allow an encounter and a successful blending of the different philological approaches [18].

On a technical level, project resources will be openly shared on the most appropriate repositories:

(1) GitHub[9] for texts annotated using the XML/TEI markup language and for the EVT software development code: this

[3] https://tei-c.org/
[4] http://www.tei-c.org/Guidelines/P5/
[5] https://www.collane.unito.it/oa/items/show/11
[6] http://pelavicino.labcd.unipi.it/
[7] http://web.unibas.it/bup/evt2/pde/
[8] http://dh.uni-wuppertal.de/test-evt-di-edition/
[9] https://github.com/

hosting will allow for a good RCS (Revision Control System) processing for all files on the basis of a widely tested and secure infrastructure;
(2) Zenodo[10] for all project documentation, for images for which rights will be granted, and for academic deliverables (conference papers, articles, workshop materials): the Zenodo platform, in addition to being exceptionally stable (it is guaranteed to preserve the data for a period of 25 years from the time when it should eventually be decided to decommission it, so to allow an orderly migration to other platforms), assigns a DOI to each resource. This makes it possible not only to securely reference the resource in question (since it is a PID: Persistent Identifier), but also to ensure a public distribution of the edition data.
The project also presents important innovative features with regard to another methodological aspect, which is particularly relevant in view of the most recent developments of Web-based digital editions, namely the distributed edition. In this latest evolution of the Digital Scholarly Edition, at least part of the core components of the edition, textual data or images, reside on external servers, outside the one on which the main edition is published, and are the product of independent projects. The progress of the current IT infrastructure of the Web, in fact, makes it possible to start the realisation of an open ecosystem, within which resources of different types are made available to all interested parties for publication in digital editions or for processing for analysis and/or visualisation purposes.
Finally, all textual data will be made available to the academic community, again by means of the Zenodo platform, according to the FAIR principles: Findable, Accessible, Interoperable and Reusable[11]. As discussed in the preceding paragraphs, all data for which rights are available will be:

- Findable: easily found on the web through the use of metadata, particularly PIDs;
- Accessible: directly downloadable or accessible on open repositories, in unnegotiated access mode;
- Interoperable: particularly with regard to textual data, all documents will be encoded making use of the XML/TEI standard, so as to ensure compatibility with other projects interested in their use;
- Reusable: distribution licences, combined with the technical aspect of interoperability, will make such data fully reusable in any kind of digital project or initiative.

4. PROJECT GOALS: INNOVATION AND INCLUSIVITY

The main goals of the project are as follows:
(1) produce and publish a series of diplomatic editions, accompanied by digital facsimiles, of the most important witnesses of the *Leges Langobardorum* textual tradition;
(2) further develop the EVT software so that it is fit for the project purposes, with particular regard to the integrated edition functionality and the accessibility settings (see below);
(3) produce and publish a critical edition of the *Leges Langobardorum*" on the basis of all available witnesses and witness fragments, with particular regard to the texts published independently as diplomatic editions;
(4) use the integrated edition feature of EVT to bring together the critical edition and the diplomatic editions, with the goal of a seamless navigation from the former to the latter;
(5) organise the dissemination of all the edition materials (those the project has rights upon) in such a way that it will be possible both to reuse them according to the FAIR principles and maximise the global impact of the project (see below);
(6) since the new edition would be the result of a research work carried out in the digital paradigm and published in a digital form, the ultimate goal is that of "disclosing" the Lombard world - strictly linked to past and present Italy within the wider European environment - to audiences beyond academia (e.g. tourism, museums and, more generally, the GLAM sector) with the aim of awakening the users to their own social memory, historical legacy and cultural heritage;
(7) Special attention will be paid to the topic of inclusion: the proponents are aware of the fact that the digital means can in itself be more inclusive (for instance, it allows remote fruition, thus facilitating, among others, people with walking impairments), but also alienating due to digital divide or the lack of specific software features that allow access to specific categories of users, e.g. visually or hearing impaired people. EVT will be equipped with accessibility settings, i.e. features that let people with disabilities customise a device for their own needs. This kind of customization can be a life-changer for people with special needs: low vision, hearing difficulties, motor control issues, auditory-processing issues, expressive language disorder, nonverbal-communication issues, as well as attentional needs.

[10] https://zenodo.org/
[11] https://www.go-fair.org/fair-principles/

REFERENCES

[1] Andrews, Tara. 'The Third Way. Philology and Critical Edition in the Digital Age'. *The Journal of the European Society for Textual Scholarship, Variants* 10 (2013): 61–76.

[2] Azzara, Claudio, and Stefano Gasparri, eds. *Le leggi dei Longobardi: storia, memoria e diritto di un popolo germanico*. Altomedioevo. Roma: Viella, 2005.

[3] Beyerle, Franz. *Die Gesetze der Langobarden. Germanenrechte. Texte und Übersetzungen 3*. Weimar: H. Böhaus, 1947.

[4] Bluhme, Friedrich, and Alfred Edwin Boretius. *Leges Langobardorum. Monumenta Germaniae Historica: Leges, t. 4*. Hannoverae: Hahnianus, 1868.

[5] Boschetti, Federico, Riccardo Del Gratta, Monica Monachini, Maria Buzzoni, Paolo Monella, and Roberto Rosselli Del Turco. 'Tea for Two': The Archive of the Italian Latinity of the Middle Ages Meets the CLARIN Infrastructure', 37–46, 2020.

[6] Buzzoni, Marina, and Roberto Rosselli Del Turco. 'Verso un'edizione digitale dell'Editto di Rotari'. In *I Longobardi in Italia: lingua e cultura. Atti del XV Seminario Avanzato in Filologia Germanica*, 37–85. Alessandria: Edizioni Dell'Orso, 2015.

[7] Dold, Alban. *Zum Langobardengesetz: neue Bruchstücke der ältesten Handschrift des Edictus Rothari*. Weimar: Böhlau, 1940.

[8] Dold, Alban. *Zur ältesten Handschrift des Edictus Rothari*. Stuttgart: W. Kohlhammer, 1955.

[9] Fischer, Franz. 'Digital Classical Philology and the Critical Apparatus'. In *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 203–20. Berlin, Boston: De Gruyter Saur, 2019.

[10] Fischer, Franz. 'Digital Corpora and Scholarly Editions of Latin Texts: Features and Requirements of Textual Criticism'. *Speculum* 92, no. S1 (2 October 2017): 265–87.

[11] Fischer, Franz. 'Representing the Critical Text'. In *Handbook of Stemmatology. History, Methodology, Digital Approaches*, edited by Philipp Roelli and Aidan Conti, 405–27. Berlin, Boston: De Gruyter, 2020.

[12] Fobelli, Maria Luigia. 'Codici miniati dell'abbazia di Cava: le Leges Langobardorum e il Beda'. *Rassegna Storica Salernitana* 11 (1989): 35–63.

[13] Leoni, Albano. *Tre glossari longobardo-latini*. Napoli: Giannini, 1981.

[14] Malaspina, Ermanno. 'Il futuro dell'edizione critica (cioè lachmanniana), più o meno digitale. Riflessioni (in)attuali. Critical (Lachmannian) Editions - a More or Less Digital Future? Reflections, New and Old'. *Storie e Linguaggi* 5, no. 1 (Special Issue: Textual Philology Facing "Liquid Modernity": Identifying Objects, Evaluating Methods, Exploiting Media) (2019): 35–60.

[15] Merkel, Paul Johannes, and Friedrich Karl Von Savigny. *Die Geschichte des Langobardenrechts: Eine Abhandlung von Johannes Merkel Als Beitrag zu Savignys Geschichte des Römischen Rechts im Mittelalter*. Berlin: W. Hertz, 1850. https://catalog.hathitrust.org/Record/010463804.

[16] Molinari, Maria Vittoria. 'Sul codice vercellese delle leggi longobarde'. In *Vercelli tra Oriente ed Occidente, tra tarda antichità e Medioevo*, edited by Vittoria Corazza, 221–47. Alessandria: Edizioni dell'Orso, 1998.

[17] Monella, Paolo. 'L'edizione scientifica digitale: la critica del testo nella storia della tradizione'. In *Textual Philology Facing 'Liquid Modernity': Identifying Objects, Evaluating Methods, Exploiting Media. Storie e Linguaggi. Rivista di Studi Umanistici*. libreriauniversitaria.it edizioni, 2019.

[18] Nichols, Stephen G. 'Introduction: Philology in a Manuscript Culture'. *Speculum* 65, no. 1 (1990): 1–10.

[19] Pierazzo, Elena. 'Digital Documentary Editions and the Others'. *Scholarly Editing: The Annual of the Association for Documentary Editing* 35, no. 23 (2014).

[20] Rosselli Del Turco, Roberto. 'Designing an Advanced Software Tool for Digital Scholarly Editions: The Inception and Development of EVT (Edition Visualization Technology)'. *Textual Cultures* 12, no. 2 (29 August 2019): 91–111. https://doi.org/10.14434/textual.v12i2.27690.

[21] Rosselli Del Turco, Roberto. *The Digital Vercelli Book. A Facsimile Edition of Vercelli, Biblioteca Capitolare, CXVII*. Collane@unito.It. Università di Torino, 2017.

[22] Rotili, Mario. *L'arte a Napoli dal VI al XIII secolo.* Napoli: Società editrice napoletana, 1978.

[23] Salvatori, Enrica, Edilio Riccardini, Roberto Rosselli del Turco, Laura Balletto, Chiara Alzetta, Chiara Di Pietro, Chiara Mannari, Raffaele Masotti, and Alessio Miaschi. *Codice Pelavicino. Edizione digitale*, 2020. https://doi.org/10.13131/978-88-944430-2-8.

[24] Venuti, Martina. 'Il manoscritto Ambrosiano B 36 Inf. testimone del Liber glossarum'. *Histoire Épistémologie Langage* 36, no. 1 (2014): 15–28.

[25] Venuti, Martina. 'L'apparato digitale di Virgilio'. In *Nuovi archivi e mezzi d'analisi per i testi poetici: i lavori del progetto Musisque Deoque, Venezia, 21-23 giugno 2010*, edited by Paolo Mastandrea and Linda Spinazzè, 29–34. Amsterdam: Adolf M. Hakkert Editore, 2011.

Un *corpus* online della letteratura secondaria (1872- 1890) del Verismo italiano

Denise Bruno[1], Giuseppe Canzoneri[2], Antonio Di Silvestro[3],
Daria Spampinato[4], Alessandro Zammataro[5]
[1] Università di Catania, Italia – denise.bruno@phd.unict.it
[2] Università di Catania, Italia – giuseppe.canzoneri@unict.it
[3] Università di Catania, Italia – antonio.disilvestro@unict.it
[4] CNR Istituto di Scienze e Tecnologie della Cognizione, Catania, Italia – daria.spampinato@cnr.it
[5] Università di Catania, Italia – alessandro.zammataro@unict.it

ABSTRACT

Il contributo illustra le prime fasi del progetto COVerLeSS (*Corpus* Online del Verismo tra Letteratura, Storia e Società) che mira a conservare, valorizzare e analizzare, in un ambiente web integrato e open access, un archivio interrogabile della letteratura secondaria (recensioni, testi giornalistici e saggistici) relativa alla produzione letteraria del Verismo italiano. Il progetto non si limita però alla costruzione di un mero archivio statico, mirando semmai all'utilizzo dei dati acquisiti e indicizzati al fine di offrire una prospettiva di lettura diacronica dell'evoluzione del lessico del Verismo. Il vocabolario *Verbum*, che consente una ricerca delle forme notevoli del lessico "standard" della corrente verista, e la timeline *Ver-in-time*, che intende collegare la letteratura secondaria ospitata sul portale ai testi primari del Verismo disponibili in diverse biblioteche digitali, costituiscono gli strumenti principali attraverso cui condurre un'analisi semantica sul piano sincronico e diacronico. Nell'ottica della fruibilità e dell'inclusività, esso si colloca nella prospettiva di una fruizione del bene culturale che riflette sia il piacere dell'accostamento al documento (riprodotto nella sua veste originaria), sia la curiosità di "navigare" all'interno del documento rappresentato in formati testuali accessibili all'utente.

PAROLE CHIAVE

Secondary literature; Verism; Lexicography; Digital philology; Digital scholarly publishing.

1. INTRODUZIONE

Il progetto fornisce un'immagine "diversa" e innovativa del Verismo, grazie al suo riverbero nei testi giornalistici e saggistici del tempo e al modo in cui essi contribuirono a costruire un'immagine dell'Italia meridionale post-unitaria a livello sociale, culturale ed economico. Sul piano della conservazione, la necessità della creazione di questo *corpus* nasce dalla forte dispersione e frammentazione bibliografica delle fonti di letteratura secondaria, quasi mai fruibili a testo pieno. D'altronde, le caratteristiche dell'archivio fanno di esso un esperimento pilota per la costruzione di raccolte della letteratura secondaria, tuttora assenti dai repertori digitali per lo studio della letteratura italiana [1, 7, 10].

Per il conseguimento degli obiettivi si stanno seguendo tre macro-fasi, ciascuna dotata di una propria metodologia: 1) conservazione; 2) valorizzazione e fruizione; 3) analisi. La prima, svolta in accordo con le Linee Guida del Piano Nazionale di Digitalizzazione, si articola in ulteriori tre fasi:

Acquisizione e metadatazione. L'acquisizione e lo stoccaggio delle immagini avvengono mediante l'applicazione di tecnologie IIIF per la gestione e la visualizzazione. La metadatazione delle fonti digitalizzate consiste nella descrizione dei supporti e dei formati secondo gli schemi standard maggiormente riconosciuti (Dublin Core, MAG, XML/TEI) e integrati in formati METS per l'interscambio.

Trascrizione e codifica. Si sta procedendo alla trascrizione del *corpus* mediante metodologie e tecnologie di OCR (come Tesseract ed eScriptorium), con la metadazione automatica del processo secondo lo standard XML-ALTO (*Analyzed Layout and Text Object*) [8]. Il testo così acquisito viene annotato da un punto di vista strutturale, filologico, linguistico, lessicale e semantico attraverso lo schema di codifica XML/TEI.

Gestione e organizzazione. L'organizzazione dei dati, per quanto riguarda le immagini, si basa sul protocollo IIIF, mentre per i metadati, trascrizione e annotazione, si appoggia su un'architettura di Database XML (eXistDB), con lo sviluppo di un'ontologia formale (RDF-XML, OWL) che estende la codifica XML/TEI, per quanto riguarda l'onomastica, la toponomastica e i titoli di opere e riviste, anche in modalità *Linked Open Data* (LOD) [6].

Questa fase consentirà di approntare, per la prima volta, la trascrizione dei testi delle riviste rendendola fruibile agli utenti mediante l'interfaccia web.

La seconda fase prevede lo sviluppo del sistema di gestione, indicizzazione e il *Search Engine* progettati e implementati

secondo le buone pratiche di ingegneria del software e dei modelli formali per la gestione di documenti testuali [5]. In un'ottica di *long term preservation*, in linea con i principi FAIR[1] [23] e TRUST[2] [14], si depositeranno i dati in infrastrutture di ricerca quali CLARIN e/o DARIAH.
L'ultima fase prevede lo sviluppo di strumenti per un'analisi dei dati attraverso alcuni algoritmi di *Natural Language Processing* e *Text Mining*.

2. IL CORPUS RIVER: MOTIVAZIONI STORICO-CRITICHE E ORGANIZZAZIONE

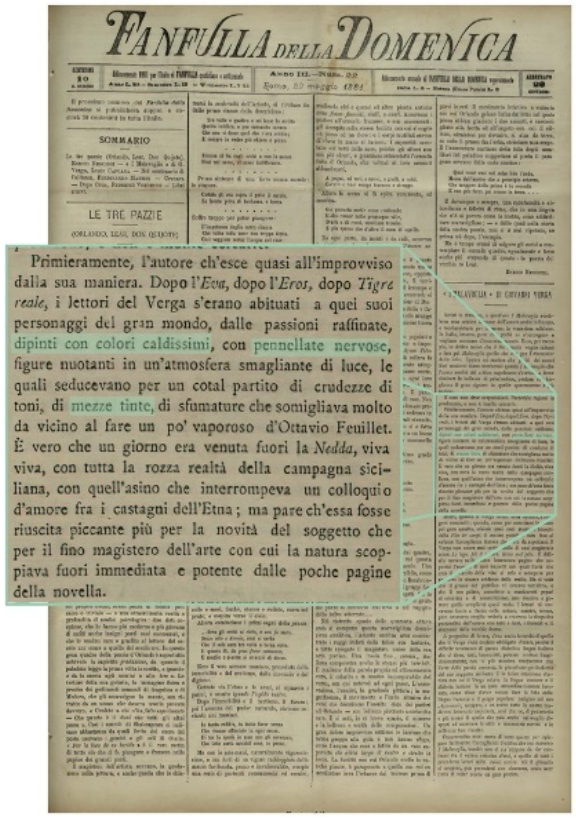
Il ventennio circa di cultura letteraria italiana che, ai fini della presente ricerca, si è scelto di comprendere tra i due estremi cronologici del 1872 e del 1890, si caratterizza per un dibattito straordinariamente partecipato da parte di letterati, scrittori e giornalisti, che si confrontano e riflettono [12], forse per la prima volta su scala nazionale, intorno ad aspetti basilari di tipo politico-economico e antropologico di una società ancora disomogenea e in via di formazione. Il tema centrale, non solo letterario, è la fedeltà al principio del "vero" in arte, alla rappresentazione della realtà nella sua forma più cruda e "naturale", che si esprime soprattutto nel "coraggio" dell'artista di scandagliare i temi scabrosi della sofferenza umana nei suoi aspetti più degradanti, quelli della povertà materiale, psicologica e morale.
Nel 1871, in prossimità dell'inizio della *timeline* contemplata dal progetto, iniziava la storia del "romanzo naturalista" con il ciclo dei Rougon-Macquart di Émile Zola [16]: un'analisi della società francese ritratta nei suoi ambienti reali − dai mercati ai sobborghi operai, dai cenacoli degli artisti alle miniere e alle campagne − e nelle sue dinamiche quotidiane. Subito dopo in Italia Luigi Capuana pubblicava la raccolta di saggi intitolata *Il teatro italiano contemporaneo* (1872), aprendo la stagione del Verismo [18]. Non a caso nel 1879 sarà ancora Capuana, con *Giacinta* (dedicato a Zola), ad inaugurare la stagione delle grandi opere veriste, anticipando, insieme a Navarro della Miraglia, che nello stesso anno pubblicava *La nana*, il Verga maggiore di *Vita dei campi* (1880), de *I Malavoglia* (1881), delle *Novelle rusticane* (1883), per arrivare fino a *Mastro Don Gesualdo* [1], che nel 1889 chiude la stagione del Verismo più produttivo e vitale.
Le opere letterarie dei veristi "maggiori" (Verga, Capuana e De Roberto), ma anche quelle che negli stessi anni andavano pubblicando i cosiddetti "minori" (Onufrio, Navarro della Miraglia, Ragusa Moleti, Chelli, Scarfoglio, Zena, Pratesi), coprono un arco cronologico lungo il quale si snoda una produzione parallela di saggi, recensioni e contributi critici, che a volte entra direttamente in dialogo con le opere letterarie (come nel caso delle recensioni), mentre in altre converge insieme a queste in un'analisi della realtà postunitaria italiana focalizzata soprattutto sulle pesanti tare economiche e culturali del Mezzogiorno, che creavano un enorme divario tra la qualità della vita delle masse contadine meridionali e quella degli abitanti del Nord Italia. Sulla base di questa riflessione preliminare, una prima parte del *corpus* di testi analizzati nell'ambito del progetto è, quindi, costituita dalla letteratura secondaria coeva alle opere del Verismo, che annovera tra i testi di maggior rilievo gli interventi critici di Francesco De Sanctis (*Il principio del realismo* del 1876 e *Studio sopra Emilio Zola* del 1879), di Felice Cameroni e di Federico De Roberto, solo per citarne alcuni. Si tratta spesso di scrittori che parteciparono in veste di critici e di cronisti al grande dibattito nazionale, in un continuo scambio di ruoli e di prospettive.
Questa produzione comprende tre tipologie di pubblicazioni: 1. i saggi sul Verismo in volume; 2. gli articoli e le recensioni pubblicate sui periodici dell'epoca e raccolti in volume dagli stessi autori; 3. le recensioni e gli articoli rimasti dispersi nelle pagine delle riviste e dei quotidiani dell'epoca [11, 17]. Le prime due categorie hanno una consistenza complessiva di circa 500 pagine e quasi equivalente è la consistenza della terza.
Un secondo blocco di testi, notevolmente più consistente di quello appena citato, è costituito da un gruppo di riviste dell'epoca, selezionate per l'impegno nella diffusione dei principi del naturalismo francese e soprattutto nella teorizzazione di una versione italiana di tale corrente letteraria: «Il Momento Letterario-Artistico-Sociale» composto da circa 880 pagine [21] e pubblicato a Palermo dal 1883 al 1885 in 72 numeri; «La Fronda», 90 pagine ca. [20], che pur pubblicata a Firenze per soli 7 numeri, tra gennaio e febbraio del 1880, ospitò articoli, recensioni e opere dei maggiori letterati italiani, tra i quali Capuana e Verga; «La Farfalla», che, dopo una prima parentesi cagliaritana (1876-1877), venne trasferita dal fondatore Angelo Sommaruga, a Milano, dove venne pubblicata fino al 1883 per 43 numeri totali (520 pagine ca.) [2]. Quest'ultima rivista si propose come una delle voci più attive, e spesso polemiche, della Scapigliatura milanese, centro di ritrovo e di confronto tra i suoi esponenti e i veristi meridionali. L'ultima delle riviste facenti parte del *corpus* è la «Rassegna Settimanale di politica, scienze, lettere ed arti», fondata da Leopoldo Franchetti e da Sidney Sonnino e pubblicata per 213 numeri dal 1878 al 1882 (4260 pagine ca.)[3], rivista che riuniva in un progetto finalmente unitario il

[1] *Findability, Accessibility, Interoperability, Reusability.*
[2] *Transparency, Responsibility, User focus, Sustainability, Technology.*
[3] https://rassegnasettimanale.animi.it/

dibattito sui temi politici, economici, sociali e letterari, dedicando particolare attenzione alle problematiche delle popolazioni del Mezzogiorno d'Italia [15].

3. *VERBUM*

Dopo la fase di acquisizione dei testi secondari del Verismo – mediante metodologie e tecnologie di OCR – si sta conducendo l'annotazione dei testi tramite lo schema di codifica XML/TEI, da un punto di vista strutturale, filologico, linguistico, lessicale e semantico. In particolare, nel *teiHeader* vengono indicati tutti i metadati, dai responsabili della digitalizzazione, della codifica e delle revisioni alla descrizione delle fonti bibliografiche; nel corpo del documento, invece, attraverso i tag `<div>`, `<p>`, `<head>` vengono codificati gli elementi strutturali, e con i tag `<persName>`, `<placeName>`, `<orgName>`, `<title>`, `<seg>`, `<term>`, `<distinct>`, `<interp>` vengono segnalati i nomi dei personaggi, dei luoghi e delle organizzazioni, i titoli delle opere citate e le annotazioni semantiche. Verrà poi implementata, con le tecnologie del Semantic Web, un'ontologia formale delle entità nominate e delle informazioni bibliografiche contenute nei file XML/TEI e che colleghi tutto il *corpus* con i repository online in modalità LOD. I lemmi, sintagmi e le strutture fraseologiche, annotati opportunamente con i tag lessicografici, formano il vocabolario *Verbum*, che permetterà di indagare i fenomeni linguistici, lessicali e semantici più rilevanti del *corpus*, con l'obiettivo di rappresentare un lemmario di base, ricercabile per forme notevoli e per varianti inerenti a un vocabolario "ideale" della corrente verista. Dunque, tale vocabolario si fonda sulla ricorsività di lemmi, di combinazioni sintagmatiche (ad es. lemma base + un attributo), o di eventuali espressioni fraseologiche, che verranno appositamente marcate per valorizzare le particolari accezioni semantiche del lemma base. Per la marcatura dei lessemi si è scelto di servirsi del tag `<term>`, da specificare con l'attributo `@type` per individuare le varie tipologie di lessico o le combinazioni sintagmatiche e fraseologiche funzionali a *Verbum*. Un esempio della marcatura lessicografica a cui sono sottoposti i testi si può desumere dalla recensione di Filippo Filippi sulla rivista «La Perseveranza» del 2 ottobre 1880, poco dopo l'uscita della raccolta *Vita dei campi*. Filippi attinge al campo semantico della vista e delle arti visive, utilizzando espressioni come *colore locale*, *osservazione continua*, *varietà dei contorni e dei colori*, *scolpiti in bronzo*. Quest'ultima espressione sarà più volte ripresa dai critici, tra tutti Capuana, il quale sottolineava (nella recensione a *Vita dei campi* apparsa sul «Corriere della Sera» del 20-21 settembre 1880) come il «bronzo della lingua letteraria» fosse calato entro la «forma sempre fresca» del dialetto. Frequenti sono, inoltre, forme verbali quali *ritratto* e *vedere*, quest'ultimo nell'accezione del visualizzare nell'immaginazione del lettore i luoghi descritti. Un'altra recensione di particolare interesse lessicale è quella di Francesco Torraca sulla rivista «Il Diritto» del 9 maggio 1881, in cui spiccano lessemi ed espressioni che appaiono come varianti di un lessico di base paradigmatico, introdotto da Verga soprattutto nei suoi pochi testi "programmatici" o nella corrispondenza privata. Esempi ne sono, nella recensione di Capuana a *I Malavoglia* edita sul «Fanfulla della domenica» del 29 maggio 1881, espressioni come *dipinte con colori caldissimi, pennellate, minute particolarità, gran quadro, schizzò quei stupendi bozzetti, crudezze di toni, di mezze tinte, di sfumature* – con ripresa delle *mezze tinte dei mezzi sentimenti* che Verga aveva utilizzato nella Prefazione a *I Malavoglia* (vd. Fig. 1)[4].


FANFULLA DELLA DOMENICA

SOMMARIO

LE TRE PAZZIE

Primieramente, l'autore ch'esce quasi all'improvviso dalla sua maniera. Dopo l'*Eva*, dopo l'*Eros*, dopo *Tigre reale*, i lettori del Verga s'erano abituati a quei suoi personaggi del gran mondo, dalle passioni raffinate, dipinti con colori caldissimi, con pennellate nervose, figure nuotanti in un'atmosfera smagliante di luce, le quali seducevano per un cotal partito di crudezze di toni, di mezze tinte, di sfumature che somigliava molto da vicino al fare un po' vaporoso d'Ottavio Feuillet. È vero che un giorno era venuta fuori la *Nedda*, viva viva, con tutta la rozza realtà della campagna siciliana, con quell'asino che interrompeva un colloquio d'amore fra i castagni dell'Etna; ma pare ch'essa fosse riuscita piccante più per la novità del soggetto che per il fino magistero dell'arte con cui la natura scoppiava fuori immediata e potente dalle poche pagine della novella.

Figura 1. Recensione di Capuana a I Malavoglia, *«Fanfulla della domenica», 29 maggio 1881*

Un primo saggio del flusso di lavoro, eseguito su un campione di 24 recensioni codificate secondo lo standard XML/TEI, ha fornito dei risultati interessanti da un punto di vista quantitativo e qualitativo. Il segmento di *corpus* analizzato comprende 9 recensioni riguardanti la novella *Nedda* (1874), 5 la raccolta *Vita dei Campi* (1880) e 10 le *Novelle Rusticane* (1882-1883).

L'analisi dei dati è stata eseguita col tool NormaTEI[5], sviluppato dal gruppo CHROMA del CNR ISTC nell'ambito del progetto BellinInRete [4]. Questo tool permette di analizzare in modo complessivo tutti i file XML/TEI di un'edizione, e in esso è

[4] Immagine con licenza CC0 1.0 Public Domain da http://www.bncrm.beniculturali.it/it/32/biblioteca-digitale
[5] https://github.com/pierpaolosichera/NormaTEI

possibile effettuare sia ricerche rapide (per individuare eventuali refusi o effettuare normalizzazioni delle codifiche), che complesse, attraverso delle interfacce intuitive. Completano il tool un ambiente statistico e sistemi di esportazione dati (txt, xls). In questo caso NormaTEI è stato utilizzato per esemplificare un campione descrittivo di *Verbum*.
Il tag scelto per rappresentare i termini all'interno di *Verbum* è `<term>`, il quale contiene una designazione composta da una sola parola, più parole o simbolica, con la quale viene identificato il termine tecnico, che nello specifico afferisce a un'area semantica fortemente caratterizzata del linguaggio verista.
L'utilizzo del tool ha consentito di isolare – all'interno delle 24 recensioni che costituiscono il mini-*corpus* preso in esame – 671 tag `<term>`, di cui ben 170 (circa il 25%) appartengono all'area semantica delle arti visive, e in generale all'atto della narrazione come rappresentazione. Nello specifico, all'interno di questo campo semantico sono state isolate una serie di "parole dominanti" (circa il 40%), le quali presentano un numero di occorrenze piuttosto elevato: esempi ne sono *bozzetto*, *colorito*, *descrizione*, *fotografare*, *quadro* – accompagnato dal ricorrente diminutivo *quadretto* – *ritrarre*, ecc. Questi "lemmi dominanti" presentano spesso delle specificazioni (ossia espansioni consistenti in aggettivi, complementi di specificazione, di modo, ecc.), che ricorrono a volte come costanti all'interno del *corpus* e altre volte come "varianti" all'interno di specifici sintagmi; è il caso di espressioni come:

Bozzetto	bozzetto dal vero
	bozzetto siciliano
Fedele	quadro fedele
	fedele pittura
Colorito	potenza di disegno e di colorito
	colorito del paesaggio
	colorito locale
	intonazione del colorito
	vivacità di colorito
Ritrarre	ritrarre la vita
	ritrarre i dialoghi
	ritrarre sentimenti
Quadro/quadretto	quadro di costumi
	quadri veri
	quadretti locali
	quadretti della vita siciliana
	quadrettino di genere campestre
	quadri variati
Dipingere	dipinge dal vero
Colore	delineare a colori vivi i contorni
	tratteggiare coi colori più smaglianti della sua tavolozza
	colore del vero
	colori vivi
	vivezza di colore
	ricchezza del colore

Per dare un'idea della ricorsività di questa area semantica, limitata alle "parole dominanti", abbiamo sottoposto la nostra selezione di 24 recensioni a un'analisi con *Voyant Tools*[6].
Il grafico che ne risulta (vd. Fig. 2) mette in evidenza le parole "pilota" già individuate con NormaTEI. Inoltre, i "picchi" delle frequenze relative (certamente ascrivibili al numero ridotto dei testi qui utilizzati), forniscono una prima, ancor grezza ma promettente, rappresentazione dell'area semantica scelta, che insieme alle altre man mano ricostruite con lo spoglio progressivo dei testi, offrirà una ricca cartografia lessicale e semantica di un eccezionale fenomeno del nostro post-risorgimento letterario quale è stato il Verismo.

[6] https://voyant-tools.org/

Figura 2. Andamenti delle parole "pilota" ricavati con Voyant Tools

La Figura 3, per esempio, restituisce graficamente la situazione delle occorrenze del lessico afferente all'area semantica delle arti visive, prendendo in considerazione il numero di occorrenze per ciascun lemma tra quelli dell'area semantica selezionata.

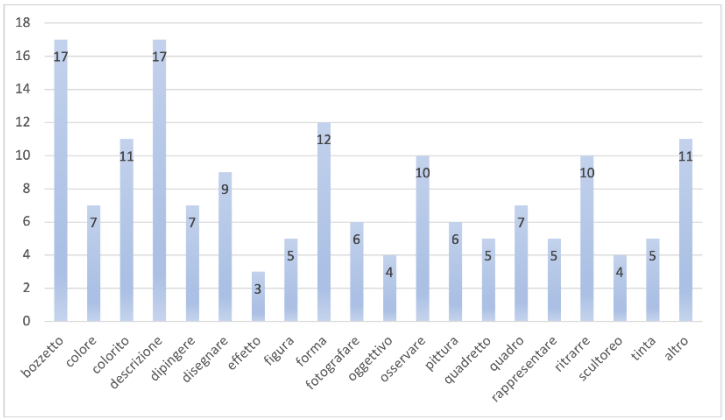


Figura 3. Lessico delle arti visive

Nella Figura 4, invece, le medesime occorrenze lessicali vengono raggruppate per affinità di significato, secondo il criterio della "parola dominante". Per esempio, le parole 'colore' (7 occorrenze) e 'colorito' (11 occorrenze) sono raccolte sotto il lemma 'colorito'. In questo grafico l'asse delle ascisse rappresenta le recensioni vagliate e facenti riferimento alle seguenti opere di Verga: *Nedda* (sigla **N**, recensioni pubblicate da giugno ad agosto 1874), *Vita dei campi* (**VC**, settembre 1880), *Novelle rusticane* (**NR**, dicembre 1882-marzo 1883); sull'asse delle ordinate, invece, è riportato il numero assoluto di occorrenze per singola recensione. In questo modo risulta evidente, dall'andamento delle linee di colore associate a ciascuna "parola dominante", l'oscillazione nell'utilizzo da parte dei recensori di Verga del lessico delle arti visive, in una progressione cronologica che va dalla prima recensione a *Nedda* del 15 giugno 1874, all'ultima recensione alle *Novelle rusticane* del 30 marzo 1883.

Oltre alle strutture sintagmatiche e fraseologiche presenti e annotate nei testi, *Verbum* sarà arricchito di sinonimi, macro e microcategorie e parole chiave, ricavati sulla base di cicli di revisione del vocabolario stesso, al fine di essere utilizzato per una indicizzazione semantica dei documenti del *corpus*, rendendoli così ricercabili nell'interfaccia di ricerca con *keywords* avanzate e accessibili nelle maschere di visualizzazione. Per fare ciò, saranno testati strumenti basati su metodologie di *Machine Learning* per la classificazione di documenti di fine Ottocento anche senza parole chiave. Alcuni documenti saranno classificati da esperti del dominio allo scopo di creare un set di apprendimento sufficientemente grande per addestrare il classificatore. In un secondo momento, si procederà alla categorizzazione di tutti i documenti del database. *Verbum,* inoltre, verrà collegato semanticamente con le opere primarie disponibili online e con altre risorse in

LOD[7], in modo da consentire interrogazioni complesse e composite su autore, anno, parole chiave, opera, titolo, abstract [3]. Inoltre, le risorse testuali oltre a essere accessibili e ricercabili, saranno rese disponibili per il download in modalità open access.

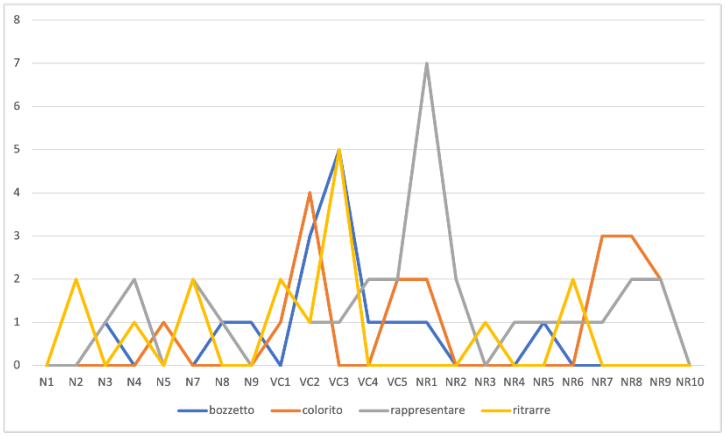


Figura 4. Lessico delle arti visive per "parole dominanti"

4. RISULTATI ATTESI

Nell'ambito del progetto, la valorizzazione e disseminazione dei risultati attesi è affidata a un portale web costituito da diversi moduli che integrano strumenti di consultazione e interrogazione avanzata. L'obiettivo è di ridefinire la fruizione e l'interpretazione della letteratura secondaria del Verismo attraverso l'impiego di strumenti di esplorazione, analisi e divulgazione del *corpus* in modo semplice ed efficace. Il primo modulo si avvale di una piattaforma web basata sul modello offerto dal software TEI Publisher[8], che consente la visualizzazione delle riproduzioni digitali delle riviste in parallelo alle loro trascrizioni. La codifica del testo è conforme allo standard XML/TEI e incorpora tag che forniscono informazioni riguardanti i metadati, le date citate all'interno dell'opera, nomi di personaggi, luoghi, incluse alcune note di carattere filologico, linguistico e lessicale (vd. Fig. 5).

Il secondo modulo sarà costituito da un motore di ricerca avanzato che consente di interrogare i testi sulla base di diversi tag utilizzati per l'annotazione del testo XML, quali autore, periodo, opere citate, nomi, luoghi e parole chiave [22].

Il terzo modulo farà uso del *Natural Language Processing* e delle librerie Javascript per fornire agli utenti una visualizzazione dei dati sul modello dello *storytelling*. Questi strumenti digitali saranno funzionali alla costruzione di una linea del tempo con sviluppo diacronico, che abbiamo denominato *Ver-in-time*.

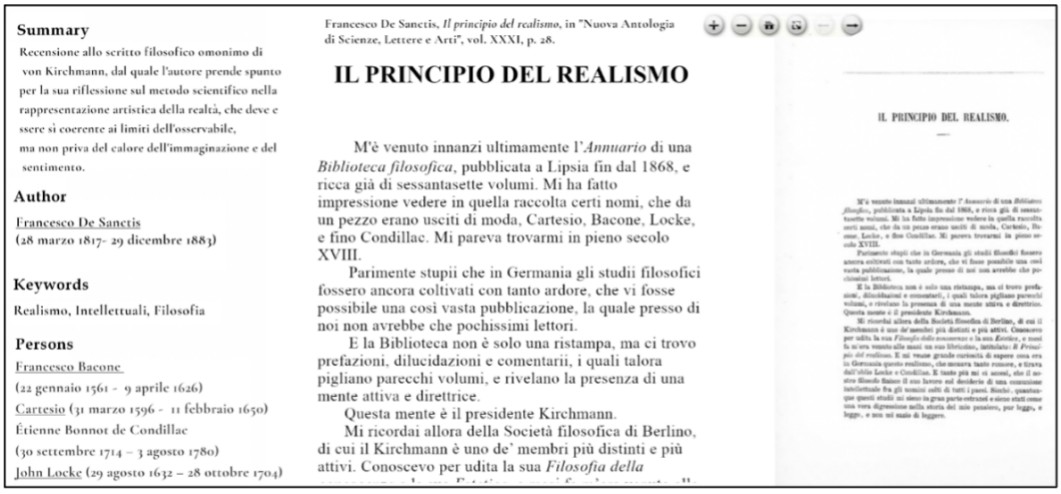


Figura 5. TEI Publisher, F. De Sanctis, «Il principio del realismo»

[7] Va segnalato, tra le risorse esterne a cui il portale intende collegarsi, VIVer (https://testi.progettoviver.it/), un *corpus* leggibile e interrogabile di testi della letteratura del Verismo esteso alle diverse espressioni regionali, in cui i testi presentano una marcatura di tipo linguistico e sintattico-fraseologico.

[8] https://teipublisher.com/exist/apps/tei-publisher-home/index.html

I tre moduli saranno integrati all'interno del portale, che si prefigge di superare i limiti entro cui sono confinati i semplici archivi di risorse digitalizzate, offrendo un ambiente interattivo per l'esplorazione, l'analisi e la divulgazione del *corpus* della letteratura secondaria relativa al Verismo [9].
La digitalizzazione integrale dei periodici selezionati, che saranno resi disponibili sul portale, non solo preserverà un bene culturale materiale a rischio di deperimento, ma lo renderà accessibile a un pubblico globale, incentivando la lettura critica e l'interazione con documenti e testimonianze letterarie e storico-culturali "di prima mano", di cui la pubblicistica periodica è un eccezionale veicolo nell'Ottocento.
Dal punto di vista analitico, il progetto incorporerà e implementerà strumenti di *Text Mining* e *Natural Language Processing* per condurre analisi quali la sentiment *analysis*, il *topic modelling* e l'estrazione di entità nominate.
Incrociare, all'interno del portale e con gli strumenti di interrogazione da esso forniti, i dati dei testi della letteratura secondaria in modo semplice e immediato può favorire uno studio più profondo e complesso delle intricate ramificazioni dei temi, delle tendenze e dei contesti storici e culturali associati al Verismo. Questi strumenti analitici non solo arricchiranno la piattaforma con nuove modalità di esplorazione dei dati, ma contribuiranno anche alla ricerca accademica fornendo nuove prospettive di analisi sul materiale testuale.
La realizzazione di questi risultati sottolinea l'impegno del progetto nel rispettare i principi FAIR per la gestione dei dati di ricerca, assicurando così che le risorse prodotte siano di ampia fruibilità e sostenibilità.
In conclusione, la piattaforma realizzata non sarà solo un ambiente digitale (mera interfaccia per sfogliare online edizioni critiche nate come cartacee), ma riprenderà il concetto che vede nell'interazione dell'utente, nella possibilità di creare percorsi "personali" a seconda dei propri interessi, il cuore della fruizione di *corpora* ed edizioni realmente digitali, in linea con quegli «ambienti, o anche infrastrutture, che raccolgono testi, servizi, interfaccia e strumenti di accesso *in usum philologorum*» [13: 113].

5. RINGRAZIAMENTI

Si ringraziano Christian D'Agata e Angelo Mario Del Grosso per il supporto nella costruzione del modello di codifica proposto e Denise Maci e Martina Corti per la codifica delle recensioni. "Finanziato dall'Unione Europea – Next Generation EU".

BIBLIOGRAFIA

[1] Briganti, Alessandra, Camilla Cattarulla, e Franco D'Intino. *I periodici letterari dell'Ottocento: indice ragionato (collaboratori e testate)*. Milano: F. Angeli, 1990.

[2] Chemello, Adriana. *«La Farfalla» di Angelo Sommaruga. Storia e indici*. Roma: Bulzoni, 1977.

[3] Cristofaro, Salvatore, Christian D'Agata, Antonio Di Silvestro, Giuseppe Palazzolo, Pierpaolo Sichera, e Daria Spampinato. «DEMOTICOn. Per un'edizione semantica dei Malavoglia». In *AIUCD2021 Book of Extended Abstracts*, (a cura di) Federico Boschetti, Angelo Mario Del Grosso, e Enrica Salvatori, 471–73. Quaderni di Umanistica Digitale, 2021. https://doi.org/10.6092/unibo/amsacta/6712.

[4] Cristofaro, Salvatore, Angelo Mario Del Grosso, Laura Mazzagufo, Pierpaolo Sichera, e Daria Spampinato. «Bellini Digital Correspondence: A Model for Making Collaborative Digital Scholarly Editions», 615–20. Agadir - Essaouira, Morocco: IEEE, 2023. https://doi.org/10.1109/CiSt56084.2023.10409920.

[5] Cristofaro, Salvatore, e Daria Spampinato. «Aspetti funzionali e implementativi del Museo epigrafico digitale EpiCUM». *Umanistica Digitale* 4, fasc. 9 (2020): 61–77. https://doi.org/10.6092/issn.2532-8816/9973.

[6] Daquino, Marilena, Francesca Giovannetti, e Francesca Tomasi. «Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini». *Umanistica Digitale* 3, fasc. 7 (2019). https://doi.org/10.6092/issn.2532-8816/9091.

[7] De Pasquale, Andrea. «The Italian national digital newspaper library». *Bibliothecae.It* 7, fasc. 2 (2018): 348–70. https://doi.org/10.6092/issn.2283-9364/8951.

[8] Del Grosso, Angelo Mario, Andrea Bellandi, Emiliano Giovannetti, Simone Marchi, e Ouafae Nahli. «Scanning is Just the Beginning: Exploiting Text and Language Technologies to Enhance the Value of Historical Manuscripts». In *IEEE 5th International Congress on Information Science and Technology*, 2018. https://doi.org/10.1109/CIST.2018.8596373.

[9] Di Silvestro, Antonio, Christian D'Agata, Giuseppe Palazzolo, e Pierpaolo Sichera. «Conservazione e fruizione di banche dati letterarie: l'archivio della poesia italiana dell'Otto/Novecento di Giuseppe Savoca». In *Atti del Convegno AIUCD2022*, 98–104, 2022.

[10] D'Orsogna, Fabio, e Giulio Palanga. «Riviste digitali e digitalizzate italiane (RIDI): a reconnaissance for the national newspaper library». *JLIS.It* 13, fasc. 1 (2022): 374–89. https://doi.org/10.4403/jlis.it-12734.

[11] Farinelli, Giuseppe. *La pubblicistica nel periodo della scapigliatura*. Milano: IPL, 1984.

[12] *I verismi regionali. Atti del Congresso Internazionale di Studi, Catania, 27-29 aprile 1992*. 2 voll. Catania: Fondazione Verga, 1996.

[13] Italia, Paola, e Francesca Tomasi. «Filologia digitale. Fra teoria, metodologia e tecnica». *Ecdotica* 11 (2014): 112–31. https://doi.org/10.7385/99218.

[14] Lin, Dawei, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, et al. «The TRUST Principles for digital repositories». *Scientific Data* 7 (2020): Article number: 144. https://doi.org/10.1038/s41597-020-0486-7.

[15] Melis, Rossana. *La bella stagione del Verga*. Catania: Biblioteca della fondazione Verga, 1990.

[16] Pellini, Pierluigi. *Naturalismo e verismo. Zola, Verga e la poetica del romanzo*. Firenze: Le Monnier Università, 2010.

[17] Rappazzo, Felice, e Giovanna Lombardo. *Giovanni Verga fra i suoi contemporanei*. Soveria Mannelli: Rubbettino, 2016.

[18] Raya, Gino. *Bibliografia di Luigi Capuana (1839-1968)*. Roma: Ciranna, 1969.

[19] Raya, Gino. *Bibliografia Verghiana (1840-1871)*. Roma: Ciranna, 1972.

[20] Romano, Cinzia. *Emanuele Navarro della Miraglia: un percorso esemplare di secondo Ottocento*. Catania: Biblioteca della Fondazione Verga, 1998.

[21] Saja, Giuseppe. *«Il Momento». Identità di una rivista di fine Ottocento con gli indici del periodico (1883-1885)*. Caltanissetta-Roma: Sciascia, 2004.

[22] Spadini, Elena, Francesca Tomasi, e Georg Vogeler, (a cura di). *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*. Norderstedt: Herstellung und Verlag, 2021.

[23] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Un progetto di edizione digitale *image-based* delle *Meraviglie d'Oriente* nel MS Cotton Vitellius A.xv

Andreea Mihaela Toma
Università degli Studi di Padova, Italia - andreeamihaelatoma@studenti.unipd.it

ABSTRACT

Le *Meraviglie d'Oriente* (*Wonders of the East*) è la traduzione antico inglese di un testo latino noto con il titolo di *De rebus in Oriente mirabilibus*. Si tratta di un trattato teratologico, nonché di un prototipo della letteratura di viaggio. In esso sono contenute descrizioni di animali bizzarri ed esseri dall'aspetto mostruoso che popolano le regioni orientali del mondo, come Babilonia e l'Egitto. La particolarità di quest'opera nel contesto manoscritto anglosassone è che tutti e tre i testi sono accompagnati da illustrazioni delle creature e dei mostri ivi descritti.

Sebbene gli editori moderni abbiano deciso di non includere le illustrazioni nelle rispettive edizioni delle *Meraviglie*, studi recenti dimostrano come queste non solo possono fornire una migliore comprensione del testo scritto, rafforzandone il significato, ma potrebbero perfino sviluppare un livello narrativo ulteriore. Un'edizione scientifica digitale *image-based* offre uno strumento che consentirà al fruitore di apprezzare l'interazione tra testo scritto e illustrazioni. Il progetto si propone di realizzare un'edizione digitale delle *Meraviglie* contenute in uno dei testimoni attraverso EVT 2; tale prodotto fornirà uno strumento utile ad uno studio di più ampio respiro del testo a tutti i suoi livelli, dando quindi giustizia alla complessità narrativa che lo caratterizza.

PAROLE CHIAVE

Cotton Vitellius A.xv; *Meraviglie d'Oriente*; illustrazioni; digital scholarly editing; EVT.

1. INTRODUZIONE

Il testo teratologico conosciuto con il titolo *De rebus in Oriente mirabilibus* arriva in Inghilterra intorno al VII secolo d.C. in una delle molte sue versioni dell'opera, nota come *Lettera di Farasmane ad Adriano*, la quale ebbe una vasta fortuna nel medioevo mediterraneo [6]. Il testo riporta la lettera che Farasmane avrebbe inviato all'imperatore Adriano, per narrargli di un suo viaggio in Oriente e delle stranezze incontrate sul suo cammino. Nel corso del tempo e già nelle prime fasi della trasmissione testuale, il *De rebus* ha abbandonato la forma epistolare, evolvendosi in un testo molto più simile ad un breve catalogo di mostri privo di una cornice narrativa [13]. Questa recensione ebbe una circolazione relativamente ampia nell'Inghilterra anglosassone, tanto che a un testimone della versione latina – Oxford, Bodleian Library, MS Bodley 614 (**MB**) – si aggiungono due testimoni della traduzione in antico inglese, traditi rispettivamente in London, British Library, MS Cotton Vitellius A.xv (**MV**) e London, British Library, MS Cotton Tiberius B.v (**MT**). Il contesto codicologico e il periodo storico in cui le *Meraviglie* si collocano influenzano la lettura e l'interpretazione dell'opera stessa. La datazione del Bodley si colloca tra il 1120 e il 1140 e qui il testo teratologico si trova insieme a un calendario e a un trattato di astronomia. Per quel concerne il Tiberius, realizzato probabilmente nel XI secolo, anch'esso è caratterizzato da una scelta testuale rivolta verso contenuti di natura pseudo-scientifica ed ecclesiastica. Tanto a **MB** quanto a **MT** viene quindi attribuito un certo grado di "scientificità". Diversi studi si sono focalizzati sul rapporto tra **MT** e **MB**, giungendo alla conclusione che molto probabilmente, sulla base del testo in latino che è praticamente identico in entrambi i manoscritti, **MT** abbia funto da antigrafo per **MB** [3]. Questa ipotesi è ulteriormente corroborata dall'apparato iconografico che accompagna le descrizioni dei mostri, molto simile in entrambi i codici. **MV**, il più antico dei tre testimoni, è stato realizzato presumibilmente intorno all'anno 1000; benché sia ipotizzato che le versioni delle *Meraviglie* antico inglesi discendano tutte da un antenato comune [5], questo manoscritto presenta una variazione testuale molto più marcata rispetto agli altri due codici. Inoltre, la collocazione delle *Meraviglie* nel contesto del manoscritto è assai particolare: esso, infatti, alterna materiale in prosa e opere poetiche, così come testi di tipo religioso accanto a opere derivanti dalla classicità greca, fino a includere *Beowulf*. Uno dei primi studiosi a essersi occupato del manoscritto è stato Kenneth Sisam, che ne ha individuato un'unitarietà tematica incentrata sul meraviglioso e sul mostruoso [12]. Sulla scia di questa teoria, Andy Orchard ha approfondito l'argomento nella sua analisi dell'elemento teratologico in questo testimone, proponendo un'edizione di alcuni testi da lui considerati, tra cui le *Meraviglie* [9]. A differenza delle versioni continentali originate dalla *Lettera di Farasmane*, tutti e tre i manoscritti insulari sono arricchiti con le illustrazioni di alcune delle creature mostruose ivi descritte e questa particolare caratteristica costituisce il punto di partenza del presente lavoro di ricerca.

2. RAPPORTO TRA TESTO E IMMAGINE

Ogni variante di un testo è portatrice di significato e prevede il riconoscimento della pluralità come tratto caratteristico della cultura medievale, in quanto "l'écriture médievale ne produit pas des variantes, elle est variance" [1]. Pertanto, questo progetto si pone come obiettivo quello di valorizzare **MV**, che, come accennato in precedenza, costituisce una redazione a sé stante rispetto agli altri due testimoni e la sua complessità lo rende un interessante oggetto di studio. Qui le *Meraviglie* riportano un testo incompleto rispetto a **MT**, in quanto il manoscritto presenta un numero rilevante di *folia* danneggiati dall'incendio che nel 1731 ha colpito la biblioteca di Sir Rober Cotton. Per questo motivo, nelle edizioni critiche si è sempre preferito **MT** a quest'ultimo. Inoltre, i primi studiosi delle *Meraviglie* hanno dato poca rilevanza all'apparato iconografico di **MV**, definendolo come un lavoro alquanto sommario e poco raffinato nel suo insieme [14]. Ciononostante, dagli studi più recenti emerge un interesse sempre crescente per l'elemento mostruoso nel manoscritto e, nello specifico delle *Meraviglie*, per la rilevanza che l'apparato iconografico riveste nell'interpretazione del testo. Ciò che ancora manca, tuttavia, è un approccio ecdotico che consideri questo testimone nella sua pluralità di significati, i quali vengono espressi tanto dal testo scritto, quanto dalle immagini [7, 10]. Al di là delle riproduzioni in facsimile delle pagine del manoscritto [4], risulta evidente come le edizioni standard del testo [9, 11] non contemplino la possibilità di integrare l'elemento visuale; quando invece lo scopo è di spostare l'attenzione sulle immagini, il testo trascritto e/o edito manca, come se si trattasse di due livelli paralleli da analizzare separatamente. Nell'edizione più recente di **MV** [7], la presenza tanto del testo quanto delle illustrazioni non sembra rispondere a criteri ecdotici fondati: le immagini si trovano assieme alla traduzione idiomatica, sono in bianco e nero e la loro *mise en page* non è fedele a quella del manoscritto, privandole del collegamento a livello semiotico con il testo verbale.

Sulla base di queste osservazioni, risulta evidente l'importanza di offrire uno strumento che permetta di rendere conto in modo preciso e metodologicamente fondato del doppio livello di lettura e interpretazione che sottende alla realizzazione di **MV**, il quale realizza pienamente ciò che Marcello Ciccuto definisce una "stratificazione dei piani di lettura" [2: 873].

3. PER UN'EDIZIONE SCIENTIFICA DIGITALE *IMAGE-BASED*

Alla luce delle considerazioni appena esposte, lo strumento digitale si configura come la soluzione più efficace per la realizzazione di un'edizione che tenga debito conto dell'apparato verbale e di quello iconografico, in quanto permette all'utente di accedere ad un livello di interattività e metatestualità che le edizioni cartacee non sono in grado di offrire. L'edizione che si intende sviluppare nel presente progetto verrà realizzata attraverso EVT 2. La scelta di utilizzare questo software è motivata da molteplici fattori, che sono:

1) la possibilità di realizzare un'edizione *image-based*;
2) la semplicità di utilizzo;
3) l'aderenza a protocolli di codifica standard;
4) l'interattività del prodotto finale;
5) la disponibilità del prodotto in modalità *open source*.

Tutti questi aspetti rendono EVT 2 lo strumento ideale per la realizzazione di un'edizione di **MV** che preveda la presenza simultanea sullo schermo del computer della riproduzione del manoscritto e della trascrizione-edizione del testo; così progettato, il lavoro offre all'utente uno strumento interattivo che permette uno studio completo del testimone, a partire dalle caratteristiche paleografiche per arrivare a quelle paratestuali, oltre ad offrire vari livelli di lettura del testo scritto, a seconda degli interessi di ricerca.

La sezione successiva illustra più nel dettaglio il *workflow* della codifica e le caratteristiche di EVT 2 che sono state utilizzate nell'edizione.

4. CODIFICA DEL TESTO E VISUALIZZAZIONE: PRASSI, PROBLEMI E RISULTATI

Poiché l'obiettivo del progetto è l'edizione *image-based* di **MV**, la codifica del testo si basa sulla modalità "Parallel Transcription" come dettato dal protocollo TEI P5[1] per le edizioni diplomatiche e/o semidiplomatiche che includono riproduzioni in facsimile. La trascrizione diplomatica riprodurrà per quanto possibile le caratteristiche grafematiche del testo, come le varianti insulari per `<g>` (ȝ), `<r>` (ꝛ), `<s>` (ſ) e `<t>` (ꞇ), e le abbreviature, come quella per la congiunzione *and* (⁊) e per il pronome dimostrativo/ relativo *þæt* (þ); la codifica prevede la loro identificazione attraverso l'elemento `<glyph>` nella `<charDecl>`, al quale corrisponde la versione normalizzata, visibile nel livello interpretativo (vd. Fig.1).

[1] https://tei-c.org/release/doc/tei-p5-doc/en/html/ST.html#STGAre.

```
<glyph xml:id="thornbar">
  <unicodeProp name="thornbar" value="LATIN SMALL LETTER THORN WITH STROKE"/>
  <mapping type="codepoint">U+A765</mapping>
  <mapping type="diplomatic">þ</mapping>
  <mapping type="normalized">þæt</mapping>
</glyph>
```

Figura 1. Codifica dell'abbreviatura per þæt *nella* `<charDecl>`

Questa edizione si pone come ulteriore obiettivo la resa di un testo linguisticamente corretto. Dato che, come accennato in precedenza, **MV** riporta una redazione incompleta e relativamente corrotta, le parti mancanti (che si sostanziano prevalentemente in piccole porzioni di parola) sono stati ricostruite sulla base della collazione con **MT**, così come gli sporadici errori meccanici di copiatura. Questa operazione avvicina il secondo livello di codifica più a un'edizione critica che ad una interpretativa; inoltre, questo livello contiene tutti quegli interventi di standardizzazione tipici di un'edizione, come, ad esempio, la corretta separazione degli elementi lessicali. Si riporta come esempio la codifica dei primi due righi del fol. 98v, che aprono le *Meraviglie*: il primo presenta l'omissione del verbo "is", che rende la frase grammaticalmente incorretta[2], la separazione piuttosto marcata dell'iniziale maiuscola del determinante "seo" per segnalare l'inizio del testo e la divisione grafica del composto "landbuend" ("territorio") nei suoi due elementi lessicali; il secondo è caratterizzato dall'assenza del grafo `<f>` di "from" a causa della bruciatura sul margine esterno del foglio, oltre alla presenza del toponimo "Antimolime" (dat. sing.), che presenta l'iniziale minuscola. Le emendazioni, per le quali si è fatto ricorso alla collazione con **MT**, sono state codificate attraverso l'elemento `<corr>`, mentre si è fatto ricorso a `<reg>` per tutti gli interventi di standardizzazione del testo originale, codificato a sua volta con `<orig>`. Si è invece deciso di optare per l'elemento `<supplied>` per aggiungere nel testo parole assenti in **MV** e presenti in **MT** al fine di una migliore intelligibilità del testo (vd. Fig. 2).

```
<choice><orig>S eoland buend</orig><reg>Seo landbuend</reg></choice> <supplied source="MT">is</supplied> on f<g ref="rins"/>uman</l>
<choice><orig><g ref="rins"/>om</orig><reg>from</reg></choice> <g ref="A"/>n<g ref="trot"/>imolime þæm</l>
```

Figura 2. Codifica dei primi due righi delle Meraviglie

La progettazione della codifica ha riguardato anche l'individuazione di possibili soluzioni a problemi non contemplati nel contesto di edizioni diplomatiche o semidiplomatiche, ma che si rivelano fondamentali nella realizzazione di edizioni critiche. Tra queste si segnalano gli interventi per unire le parole che nel manoscritto compaiono divise dalla fine del rigo e l'inizio di quello nuovo e l'inclusione dei segni interpuntivi moderni. Per quanto riguarda il primo aspetto, si riporta un esempio tratto ancora dal primo foglio di **MV**, dove il copista ha separato la parola "acenned" ("nato") trascrivendo l'elemento lessicale *acen* nel rigo 14 e la marca morfologica di participio passato *-ned* in quello successivo. Al fine di segnalare l'unione tra questi due elementi, si è deciso di aggiungere a livello di codifica un trattino (*dash*) dopo `<acen>`; tale indicazione ricalca la prassi moderna ed è pertanto facilmente riconoscibile nella fase di lettura del livello interpretativo (vd. Fig. 3).

```
<g ref="THORN"/>æ<g ref="rins"/> beoð <g ref="wynn"/>eð<g ref="rins"/>a<g ref="slong"/> <choice><orig>acen</orig><reg>acen-</reg></choice></l>
ned on oxna micelne<g ref="slong"/><g ref="semicolon"/></l>
```

Figura 3. Codifica di "acenned"

La codifica esemplificata sopra evidenzia anche alcune scelte che si sono imposte nella decisione di inserire la punteggiatura moderna. A tal fine, si è reso necessario l'utilizzo di strategie di codifica che possano rendere possibile la visualizzazione dei segni interpuntivi aggiunti nel livello interpretativo, senza intaccare quello diplomatico ed evitando che EVT 2 li segnalasse alla stregua di interventi critici. Si è deciso quindi di ricorrere alla `<charDecl>` con la creazione di sequenze di segni interpuntivi richiamati con `<g ref>` che realizzino elementi assenti nel livello diplomatico; lo stesso criterio è stato utilizzato per l'indicazione dell'iniziale maiuscola dopo il punto (vd. Fig.4).

```
bu<g ref="rins"/>h<g ref="fullstop"/> <g ref="THORN"/>onon <g ref="slong"/><g ref="ydot"/>ndon</l>
```

Figura 4. Stralcio della codifica del rigo 18, dove "burh" ("città") in fine di frase è seguito da un punto che si rende visibile solo a livello interpretativo, così come l'iniziale maiuscola dell'avverbio "þonon" ("lì").

[2] L'intera frase, attestata ai primi tre righi, recita "Seo landbuend [is] on fruman from Antimolime þæm lande".

Le soluzioni qui illustrate hanno carattere non definitivo e potranno senz'altro essere rimpiazzate da opzioni di codifica più efficaci; tuttavia, nel contesto sperimentale di questa edizione esse risultano abbastanza soddisfacenti e rispondenti agli obiettivi del progetto.

La visualizzazione delle immagini del manoscritto, disponibili sul sito della British Library, è possibile grazie all'applicazione del protocollo IIIF supportato da EVT 2 [8]; al momento però, dati i noti problemi legati all'attacco informatico subito dal sito della biblioteca, in fase preliminare verranno utilizzate riproduzioni in .jpg. Si è inoltre deciso di sfruttare la modalità "Image-Text Linking" offerta dal software per una rapida e comoda individuazione della corrispondenza tra testo edito e facsimile.

Le schermate seguenti illustrano l'edizione del primo foglio di **MV** così progettata (vd. Figg. 5-6):

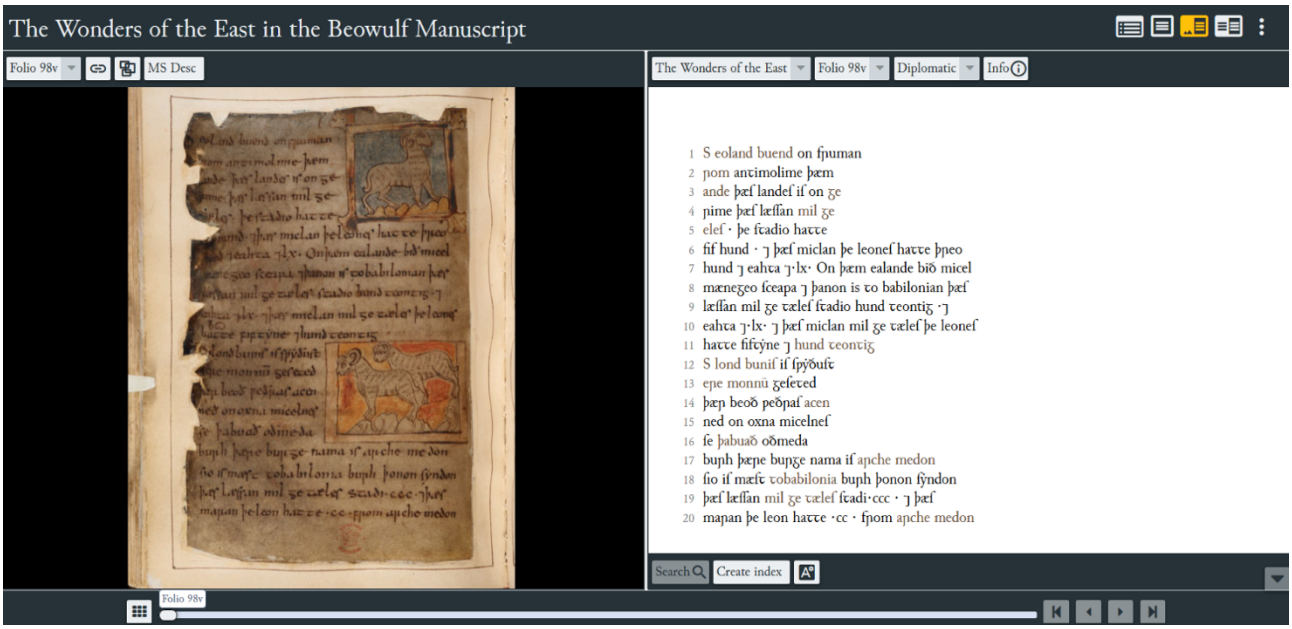


Figura 5. SDE *delle* Meraviglie *con EVT 2 (livello diplomatico)*

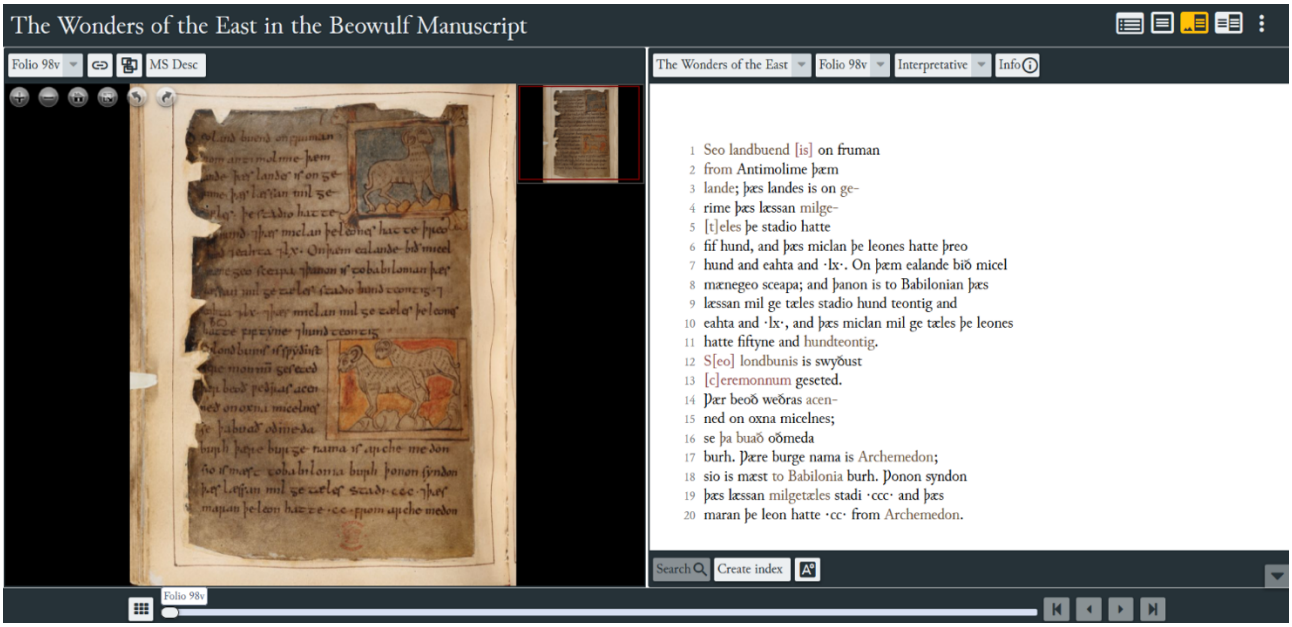


Figura 6. SDE *delle* Meraviglie *con EVT 2 (livello interpretativo)*

Un ulteriore problema riscontrato nella realizzazione della presente edizione è, oltre alla summenzionata aggiunta di parole assenti in **MV** a cui si è ovviato con l'elemento `<supplied>`, l'inserzione di intere pericopi senza le quali la comprensione del testo risulterebbe difficile. Per esempio, nel fol. 105r il rigo 8 riporta dopo "hyhst" la frase "to cynedome þone Readan Sæ and to anwalde", isolata tanto rispetto alla frase precedente quanto a quella successiva. In effetti, la traduzione sarebbe "potere e dominio sul Mar Rosso" e sembrerebbe omettere una parte iniziale, poiché è poco probabile

che si riferisca alla montagna descritta nella frase precedente. Rypins e Orchard nelle loro edizioni aggiungono da **MT** la pericope "Þær syndon gedefelice menn þa habbað him", il cui significato è "Lì vi sono popoli decenti che hanno". Per rendere questo intervento nell'edizione interpretativa, si è deciso di utilizzare il tag `<note>` (vd. Fig. 7). Il fol. 105r mette in evidenza un altro problema individuato nella codifica di questa edizione, ovvero l'allineamento del testo dell'edizione diplomatica e interpretativa con la sua collocazione sul foglio manoscritto.

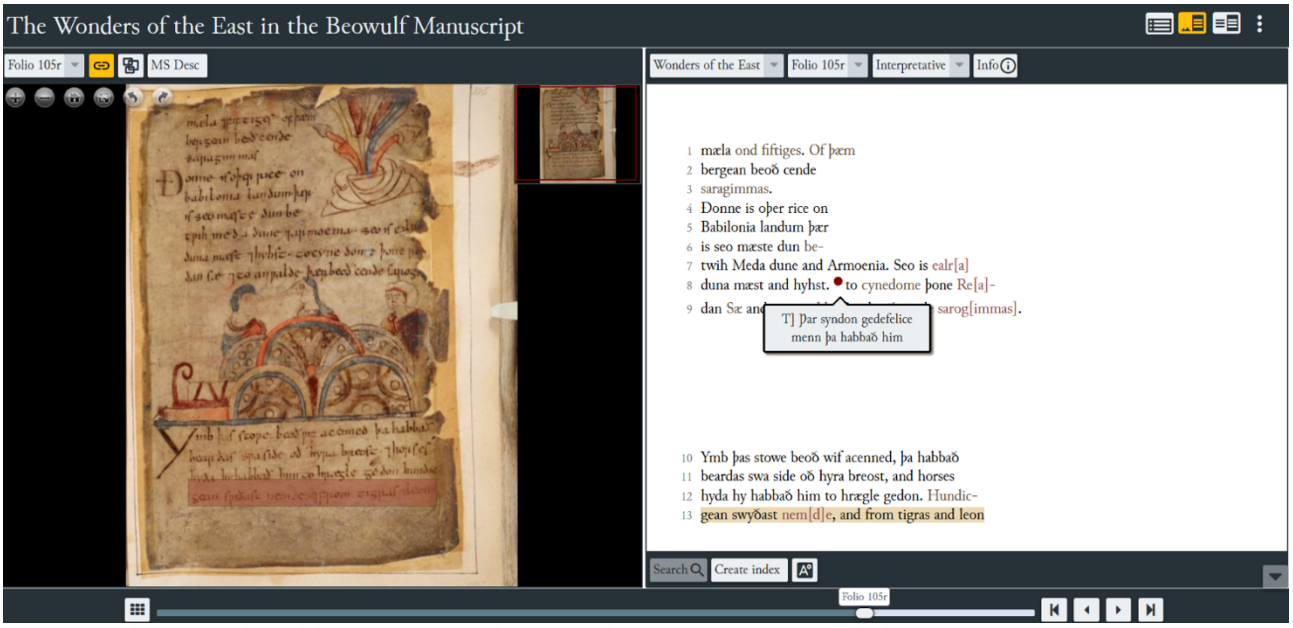


Figura 7. Fol. 105r e utilizzo del tag `<note>` per l'aggiunta di una pericope proveniente da MT

Per concludere, questo tipo di edizione si dimostra una soluzione efficace ai fini di una ricerca su più fronti di **MV** e offre uno strumento versatile e interattivo, ideale per un testo, come dimostrato in precedenza, di non facile categorizzazione ma estremamente interessante per la ricchezza semantica che esprime attraverso le parole e le immagini. Nonostante i risultati raggiunti con EVT 2 siano molto positivi, sarebbe auspicabile nel contesto del presente progetto implementare a livello di software la possibilità di poter unire le parole nel livello interpretativo, senza dover ricorrere al trattino, nonché la possibilità di evidenziare le immagini e aggiungervi eventuali informazioni per fornire una lettura più completa dei vari livelli narrativi. Nella sezione testuale, non è stata trovata una soluzione per riprodurre in maniera soddisfacente l'apparato iconografico, per cui l'edizione risulta suddivisa in sezioni separate da uno spazio vuoto.

BIBLIOGRAFIA

[1] Cerquiglini, Bernard. *Éloge de la variante: Histoire critique de la philologie*. Paris: Éditions du Seuil, 1989.

[2] Ciccuto, Marcello. «Figure dell'enciclopedia illustrata nel *De rebus in Oriente mirabilibus*». *Latomus* 52, fasc. 4 (1993): 868–74.

[3] Ford, Alun J. *Marvel and Artefact: The 'Wonders of the East' in Its Manuscript Contexts*. Leiden: Brill, 2016.

[4] James, Montague R. *Marvels of the East: A Full Reproduction of the Three Known Copies s, With Introduction and Notes*. Oxford: Roxburghe Club, 1929.

[5] Lendinara, Patrizia. «Di meraviglia in meraviglia». In *Circolazione di uomini, di idee e di testi nel Medioevo germanico: Atti del XXV Convegno dell'Associazione Italiana di Filologia Germanica*, (a cura di) Franco De Vivo, 177–229. Cassino: Edizioni dell'Università degli studi di Cassino, 2002.

[6] Lendinara, Patrizia. «Le versioni anglosassoni delle *Meraviglie d'Oriente*: varianti e variazioni». In *Il fantastico nel Medioevo di area germanica: Atti del XXXI Convegno dell'Associazione Italiana di Filologia Germanica*, (a cura di) Lucia Sinisi, 35–78. Bari: Edipuglia, 2015.

[7] Mittman, Asa Simon, e Susan M. Kim. *Inconceivable Beasts: The 'Wonders of the East' in the 'Beowulf' Manuscript*. Arizona: ACMRS, 2013.

[8] Monella, Paolo, e Roberto Rosselli Del Turco. «Extending the DSE: LOD Support and TEI/IIIF Integration in EVT». In *Atti del IX Convegno Annuale AIUCD. La Svolta Inevitabile: Sfide e Prospettive per l'Informatica Umanistica*, (a cura di) Cristina Marras, Marco Passarotti, Greta Franzini, e Eleonora Litta, 148–55. Quaderni di Umanistica Digitale, 2020. https://doi.org/10.6092/UNIBO/AMSACTA/6316.

[9] Orchard, Andy. *Pride and Prodigies: Studies in the Monsters of the 'Beowulf' Manuscript*. Cambridge: D. S. Brewer, 1995.

[10] Phillips, Peter M. «The Power of Visual Culture and the Fragility of the Text». In *Ancient Manuscripts in Digital Culture: Visualisation, Data Mining, Communication*, 30–49. Leiden: Brill, 2019. https://doi.org/10.1163/9789004399297_004.

[11] Rypins, Stanley. *Three Old English Prose Texts in MS Cotton Vitellius A. xv: 'Letter of Alexander the Great', 'Wonders of the East', 'Life of St. Christopher'*. EETS 161. London: Oxford University Press, 1924.

[12] Sisam, Kenneth. *Studies in the History of Old English Literature*. Oxford: Clarendon Press, 1953.

[13] Thomson, Simon C. *Communal Creativity in the Making of the 'Beowulf' Manuscript: Towards a History of Reception for the Nowell Codex*. Leiden: Brill, 2018.

[14] Thomson, Simon C. «The Two Artists of the Nowell Codex Wonders of the East». *SELIM. Journal of the Spanish Society for Medieval English Language and Literature* 21 (febbraio 2019): 105–54.

Una proposta di codifica in XML/MEI per testi musicali autografi di Vincenzo Bellini

Laura Mazzagufo
CNR Istituto di Scienze e Tecnologie della Cognizione, Italia - laura.mazzagufo@istc.cnr.it

ABSTRACT

Nel contributo sono descritti i criteri con cui è stata realizzata la codifica, utilizzando il vocabolario XML/MEI, di una selezione di schizzi belliniani del fondo musicale del Museo civico Belliniano di Catania, mettendo in luce le particolarità del testo musicale manoscritto e l'integrazione dei dati di interesse musicale con quelli codificati all'interno di risorse esterne.

PAROLE CHIAVE

XML/MEI; Vincenzo Bellini; music encoding; digital scholarly edition; digital textual scholarship.

1. INTRODUZIONE E OBIETTIVI

I testi, da oggetto di studio delle discipline prettamente umanistiche, sono da tempo al centro di un filone di ricerca che interessa le scienze e le tecnologie dell'informazione e che ha prodotto, oltre a nuove prospettive su strumenti e metodologie, anche interessanti interazioni sul piano dei contenuti [4]. Tale ricerca ha dato risultati particolarmente evidenti nella codifica dei testi, un settore che – contribuendo non solo alla conservazione, ma anche all'accesso, all'elaborazione digitale e al trattamento automatico dei testi – è oggi tra i più produttivi in tale ambito. Negli ultimi anni, la comunità scientifica ha rivolto la propria attenzione anche su una categoria particolare di testi, quelli musicali, avviando alcuni importanti imprese editoriali in ambito digitale[1], con le relative campagne di codifica e alcune significative riflessioni metodologiche. Il testo musicale – pur adottando un codice diverso, la notazione – condivide con quello verbale i ben noti problemi di trasmissione e di interpretazione [3], ai quali si aggiungono, nel processo di codifica, quelli di natura rappresentazionale. Alcune specificità del testo musicale, d'altro canto, hanno sollevato delle perplessità tra coloro che si occupano di *digital scholarly editing* e sono tuttora al centro di un acceso dibattito. In questo contributo si darà un breve saggio delle criticità che un testo musicale può sottoporre all'attenzione del codificatore, limitando la materia alla trattazione di alcuni casi esemplari riscontrati nella codifica di un testo autografo del compositore catanese Vincenzo Bellini. La casistica illustrata è perciò strettamente legata alle caratteristiche di una particolare tipologia testuale, ovvero un testo manoscritto, databile attorno agli anni Trenta del diciannovesimo secolo, in uno stato di abbozzo o genericamente embrionale, che prevede la compresenza di più codici (notazione musicale e annotazioni testuali) e diverse particolarità notazionali.

2. STATO DELL'ARTE E METODOLOGIA

La rappresentazione tramite tecnologie XML-based[2] di una selezione dei materiali manoscritti conservati preso il fondo musicale del Museo civico Belliniano di Catania può rappresentare un'operazione molto utile per la ricerca accademica su vita e opere del compositore, specie se si considerano gli importanti sviluppi che hanno interessato gli studi belliniani in ambito musicologico dal 2001 – anno del bicentenario della nascita di Bellini – ad oggi [9].

Al fine di garantire una codifica coerente e consentire la fruizione delle informazioni codificate in differenti modalità di presentazione, il testo musicale è stato codificato in conformità alle indicazioni delle linee guida della Music Encoding Initiative (MEI)[3]. Il vocabolario MEI offre una grande flessibilità per la codifica di documenti musicali di vario tipo e si adatta a un'ampia varietà di usi; al contempo, è evidente il debito di quest'ultimo nei confronti del più longevo e utilizzato

[1] Si citano, come esempi particolarmente significativi, la *Digitale Mozart-Edition* (DME, https://dme.mozarteum.at/) a cura della Fondazione Mozarteum di Salisburgo (ISM) in collaborazione con The Packard Humanities Institute (PHI) di Los Altos (California) e il *Beethoven Werkstatt* (https://beethovens-werkstatt.de/), frutto del lavoro collaborativo tra la Beethoven-Haus di Bonn e il Musikwissenschaftliches Seminar di Detmold/Paderborn, e finanziato dall'Akademie der Wissenschaften und der Literatur di Mainz (Magonza).

[2] La codifica dei materiali autografi musicali è stata elaborata a partire da testi già studiati criticamente, e in particolare si è fatto riferimento allo studio e alle trascrizioni di Mantica in [6] e all'edizione critica dell'opera *I Puritani* [2].

[3] Kepper, Johannes, and Roland, Perry D. *Music Encoding Initiative Guidelines*. [v. 5.0]. https://music-encoding.org/guidelines/v5/content/index.html

vocabolario della Text Encoding Initiative (TEI)[4]: le stesse linee guida della MEI ammettono il ricorso a un decisivo ed esteso "prestito" nei confronti di quelle della TEI, seguite come esempio[5]. Allo stesso tempo, l'eventuale ed auspicata integrazione tra i due vocabolari è ancora una questione aperta e molto dibattuta [12], specie nella realizzazione di edizioni scientifiche digitali [7]: lo *Special Interest Group on Music* della TEI ha avanzato alcune proposte per la realizzazione di una personalizzazione ODD in grado di incorporare estratti in MEI in un documento codificato secondo lo standard TEI, ma le informazioni rintracciabili a tal proposito sono datate e insufficienti e a tutt'oggi non è stata realizzata alcuna personalizzazione ufficiale di uno schema di codifica MEI che permetta di includere elementi appartenenti al *namespace* TEI e viceversa[6].

La documentazione fornita dalla MEI[7] è stata fondamentale per definire un iniziale modello di codifica e per consultare numerosi esempi applicati a una molteplice varietà di testi. È stata inoltre di notevole utilità la disamina dei criteri di codifica applicati in progetti di edizione di testi musicali liberamente consultabili online, e in particolare quelli che trattano repertori simili (musica manoscritta oppure abbozzi e schizzi d'autore)[8].

3. CRITERI DI CODIFICA

La selezione dei materiali codificati si è focalizzata su alcuni segmenti di testo musicale estratti dal fascicolo MM.B.36 del Museo civico Belliniano, principale testimone[9] dei cosiddetti 'studi giornalieri': si tratta di dieci fogli cartacei, autografi di Vincenzo Bellini, interessati da scrittura sul fronte e sul retro di ogni pagina (vd. Fig. 1).

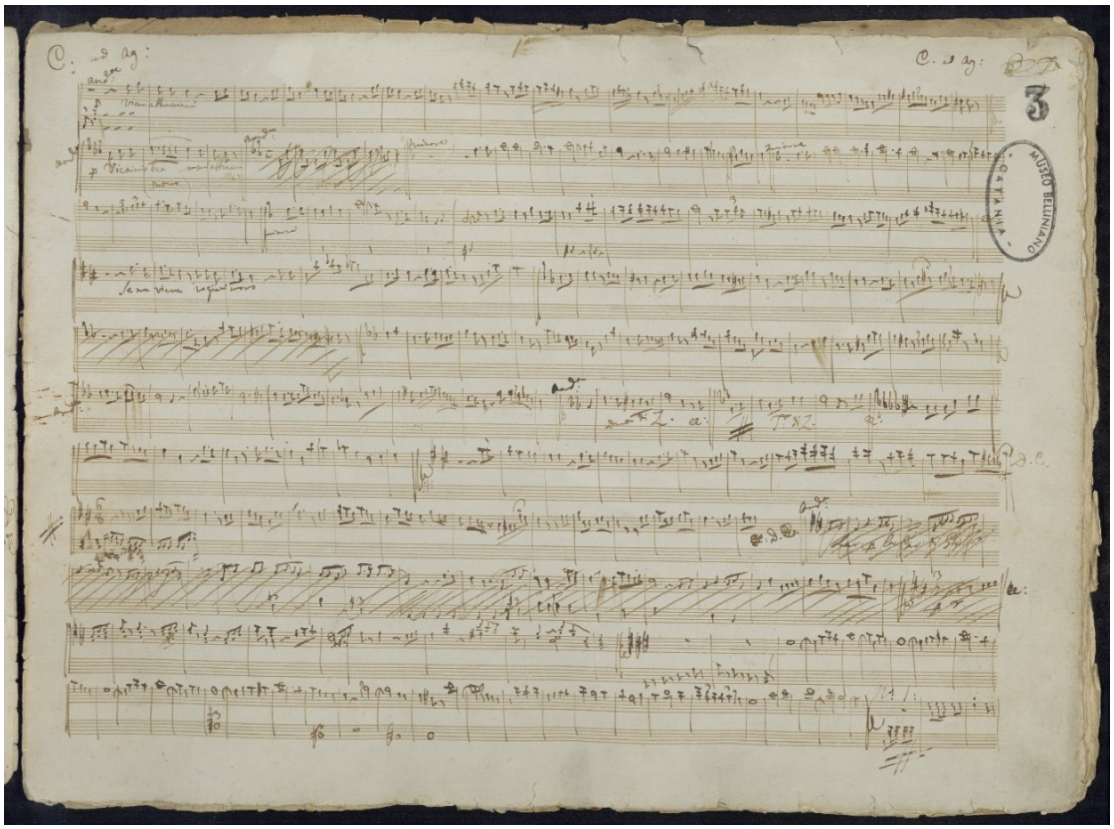


Figura 1. Riproduzione della pagina 3 (carta 2 recto*) del fascicolo MM.B.36 del Museo civico Belliniano. © Comune di Catania - Museo civico Belliniano*

[4] Text Encoding Initiative Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 2023. https://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

[5] Si veda il paragrafo *1.1.1 Acknowledgments* nell'introduzione alle linee guida MEI.

[6] A tal proposito, si segnala la Joint MEC and TEI Conference (Paderborn University, 4-8 settembre 2023) i cui atti sono stati pubblicati in [11], e in particolare [8].

[7] https://music-encoding.org/resources/tools.html. Si veda in particolare il MEI Sample Encodings repository all'indirizzo https://github.com/music-encoding/sample-encodings.

[8] Si rimanda al già citato *Beethoven Werkstatt* o il progetto *Online Chopin Variorum Edition* (OCVE, https://chopinonline.ac.uk/ocve/), un archivio interconnesso di manoscritti digitalizzati e fonti stampate relative all'opera del compositore polacco, realizzato dall'A. Mellon Foundation con l'University of Cambridge e il King's College di Londra.

[9] Perlomeno alla data odierna. Per l'individuazione di tali segmenti – particolarmente interessanti sotto il profilo melodico, genetico o storico – si è fatto riferimento a [6].

Sebbene il contenuto musicale di tale testimone – almeno in parte databile intorno ai mesi parigini dell'arco biografico del compositore – non sia direttamente correlato a una specifica opera, la sua importanza risiede tuttavia nel fatto che nei dieci fogli sono annotati circa 387 'studi', ovvero idee musicali in stato embrionale, di differente natura ed estensione, alcune delle quali sarebbero poi state utilizzate per la composizione dei *Puritani*. Gli studi sono annotati consecutivamente su un sistema composto da due pentagrammi (quello inferiore utilizzato solo saltuariamente, per annotare una struttura armonica o un accompagnamento particolare) e sono caratterizzati – già in questa fase preliminare – da una connotazione tonale ben precisa, sebbene non sempre corrispondente con quella adottata nel testo pubblicato nei *Puritani* o altrove.
Come spesso accade per schizzi e materiale manoscritto di questo genere, le chiavi e le indicazioni di tempo sono quasi sempre implicite e si registra la presenza di un gran numero di varianti sostitutive e alternative. Inoltre, la maggior parte dei fogli presenta una o due sigle a margine, generalmente sull'angolo superiore sinistro, con cui Bellini annota la natura o la destinazione degli studi presenti sul foglio. Si tratta quindi di materiali eterogenei e differenti per forma e contenuto, che annoverano non solo melodie progettate per specifiche parti vocali, ma anche passi strumentali e accompagnamenti. È stato notato [1] come nell'epistolario il riferimento a tali studi preparatori avvenga indifferentemente con il termine «mottivo» o «idea»[10]: parte di tale materiale, quindi, può esemplificare alcuni passi della corrispondenza belliniana, o può essere utile per contestualizzare – se non "dimostrare" – alcune precisazioni o dichiarazioni d'intenti che traspaiono dalle lettere del compositore ai corrispondenti più stretti.
Il documento XML/MEI[11] con cui sono stati codificati gli 'studi giornalieri' selezionati è stato impostato utilizzando l'elemento radice `<meiCorpus>`[12]: per fornire al lettore il contesto nel quale le idee melodiche degli 'studi' sono state successivamente rielaborate dal compositore, si è fatto inoltre ricorso al testo critico dell'edizione dei *Puritani* [2], e in particolare alla porzione del passo del numero operistico correlato agli 'studi' nello stadio finale (strumentazione completa). Quest'ultimo è stato codificato in un elemento `<mei>` preceduto da un ulteriore elemento `<mei>` destinato alla rappresentazione dei singoli 'studi', a loro volta raggruppati in un elemento `<group>`.

4. CASISTICA ED ESEMPI

Alcune delle principali particolarità degli schizzi belliniani in esame hanno trovato una precisa rappresentazione all'interno della codifica. La sommaria classificazione interna redatta dal compositore stesso con sigle a margine delle pagine, ad esempio, è stata registrata all'interno della sezione `<front>` di ciascuno degli 'studi' trascritti (vd. Fig. 2). Le sigle sono state sciolte nel rispettivo elemento `<expan>`, secondo l'ipotesi interpretativa proposta da Mantica [6].
I parametri generici della musica sono definiti in `<scoreDef>` – dove è possibile indicare l'armatura di chiave (che identifica la tonalità del testo musicale) e l'indicazione metrica del tempo – e l'elemento `<staffDef>` è usato per descrivere il singolo pentagramma (ad esempio il tipo di chiave utilizzata): poiché questi elementi – la chiave e l'indicazione metrica, in particolare – sono spesso lasciati impliciti negli autografi degli 'studi giornalieri', la codifica si è avvalsa dell'attributo `@clef.visible` di `<staffDef>` (il cui valore è un booleano, e «determines whether the clef is to be displayed»[13]) e dell'attributo `@meter.visible` di `<scoreDef>`, con funzione equivalente. Tramite l'elemento `<supplied>`, invece, si è dato conto dell'intervento editoriale a cura del trascrittore.
Nella codifica del testo musicale si è fatto riferimento alla sezione *11.2 Editorial Markup* delle linee guida[14]. In particolare, gli elementi `<del>` e `<add>`, mutuati dal vocabolario TEI ed eventualmente corredati dagli attributi `@resp` e `@cert`, sono stati utilizzati per registrare, rispettivamente, la presenza di una cancellatura e di un'aggiunta di testo, in combinazione con l'elemento `<subst>` quando rappresentano una variante sostitutiva. Più complessa è apparsa invece la casistica correlata alla codifica delle varianti alternative: in generale, è apparso poco opportuno avvalersi di un elemento `<rdg>` figlio di `<app>`, dal momento che la variante alternativa è attestata nel medesimo testimone[15]. Si è optato invece per una

[10] Di veda ad esempio la lettera di Vincenzo Bellini a Francesco Florimo (Puteaux, 4 ottobre 1834) in [1], p. 401-3: «[…] due pezzi composti resta ad istrumentarli e metterli bene insieme, perché ho i mottivi principali preparati ed aspetto Pepoli che mi finisca un duetto che ne spero molto».
[11] Il documento XML integrale è consultabile integralmente sul repository GitHub al seguente indirizzo: https://github.com/LauraMazzagufo/StudiGiornalieri
[12] Un elemento `<meiCorpus>` definisce «a group of related MEI documents, consisting of a header for the group, and one or more `<mei>` elements, each with its own complete header» (https://music-encoding.org/guidelines/v5/elements/meiCorpus.html). La documentazione su tutti gli altri elementi citati nel contributo può essere reperita all'indirizzo https://music-encoding.org/guidelines/v5/elements.html.
[13] https://music-encoding.org/guidelines/v5/elements/staffDef.html.
[14] Vd. nota 3.
[15] In base alle linee guida MEI, infatti, il modulo 11.1 *Critical Apparatus*, cui fanno riferimento elementi come `<app>`, `<lem>` o `<rdg>`, «describes how to encode differences between multiple exemplars of the same musical work (often referred to in MEI as 'sources')» e

soluzione meno specifica ma più coerente con quanto rappresentato nel testo, includendo ciascun gruppo di due misure che appartengono a una variante alternativa in una differente `<section>` caratterizzata dall'attributo `@type="variante_alternativa"` e messe in relazione tra loro tramite l'attributo `@target`.

```
<front>
  <div type="sigle_belliniane"><!-- sigle di Bellini -->
    <p n="1" xml:id="sigla1" hand="#VB_brownInk1" resp="#CBM">
      <choice>
        <abbr>C<am>:</am> ed ag<am>.</am></abbr>
        <expan>C<ex>anta</ex>to ed ag<ex>ito</ex></expan>
      </choice>
    </p>
    <p n="2" xml:id="sigla2" hand="#VB_brownInk2" cert="low" resp="#CBM">
      <choice>
        <abbr>C<am>.</am> ed ag<am>.</am></abbr>
        <expan>C<ex>anta</ex>to ed ag<ex>ito</ex></expan>
      </choice>
    </p>
    <p n="3" xml:id="sigla3" hand="#VB_brownInk1" resp="#CBM">
      <del rend="strikethrough">
        <choice>
          <abbr>P T</abbr>
          <expan>P<ex>ezzi</ex> T<ex>eatrali</ex></expan>
        </choice>
      </del>
    </p>
  </div>
  <div type="paratesto">
    <stamp type="catalogue">
      <locus n="3">3</locus>
      <corpName ref="#museo_belliniano">Museo Civico Belliniano </corpName>
    </stamp>
  </div>
</front>
```

Figura 2. Sezione `<front>` della codifica di uno degli 'studi giornalieri', con particolare riferimento alla codifica delle sigle belliniane che si leggono nella pagina 3 del fascicolo MM.B.36 (si veda la figura 1, margine in alto).

5. QUESTIONI APERTE E PROSPETTIVE FUTURE

Il vocabolario MEI mette a disposizione gli elementi del modulo *11.3.1 Encoding Genetic States* per descrivere le variazioni e gli sviluppi all'interno di ciascuno stadio genetico del testo: si è scelto di fare quindi riferimento al tagset `<genDesc>` per dare conto delle relazioni di consequenzialità e dei legami di vario genere tra uno 'studio' e l'altro. Tuttavia, questa configurazione (che comprende una descrizione di ciascun *genetic state* in `<genState>` con il relativo riferimento tramite l'attributo `@state`) è proposta nelle linee guida principalmente per designare specifiche campagne correttorie, e difatti `@state` può essere usato come attributo solo di elementi che definiscono un preciso intervento editoriale da parte dell'autore. Si è quindi ipotizzato l'uso dell'elemento milestone `<relation/>`, il quale «describes a relationship or linkage amongst entities»[16] e utilizza gli attributi `@plist`, `@target` e `@rel` per puntare al corrispondente elemento `<item>` di `<manifestationList>` che, nel `<meiHead>` del corpus, rappresenta e descrive ciascuno degli 'studi'.
Secondariamente, si è cercato di mettere in luce la connessione tra gli schizzi (e quindi gli stati genetici più antichi) e la realizzazione cronologicamente più tarda del testo musicale, esemplificando il caso tratto dall'opera *I Puritani*, e in particolare in alcuni spunti melodici confluiti nel tempo di attacco del duetto di Arturo ed Elvira, situato nel terzo atto dell'opera (n. 9, "Nel mirarti un solo istante")[17]. Il collegamento tra 'studi giornalieri' e partitura dei *Puritani* è evidenziato tramite l'uso dei principali elementi del modulo 11 della MEI, ovvero `<app>`, `<lem>` e `<rdg>`, che hanno fornito una possibile strategia per registrare una porzione di testo – in questo caso, ai fini del confronto, è stato utilizzato come testo di riferimento quello dell'edizione critica, registrato in `<lem>`) – e i relativi passi negli 'studi giornalieri' di qualche interesse nella genesi del testo. Per fare riferimento a quest'ultimi sono stati presi in considerazione:

- l'attributo `@target`, il cui valore punta a un `@xml:id` di un elemento XML/MEI degli 'studi' codificato nello stesso documento;

non singole lezioni alternative reperite sulla medesima fonte. Tuttavia, soluzioni di questo genere sono attestate in vari studi sul markup per la filologia d'autore, specie per la codifica in XML/TEI di varianti genetiche (si veda, ad esempio, [10]).
[16] https://music-encoding.org/guidelines/v5/elements/relation.html. Ancora una volta, il modello seguito è quello del rispettivo elemento TEI (https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-relation.html).
[17] L'interessante percorso evolutivo di una parte dei motivi melodici o di accompagnamento – sia all'interno di famiglie di 'studi' sia nel riutilizzo del materiale in partitura – è descritto in [6] ed è a questo studio che si fa riferimento per le informazioni utilizzate nella codifica.

- l'attributo `@targettype`[18], che identifica il tipo di relazione che si instaura tra la porzione di testo indicata in `<lem>` e quella identificata in `@target`;
- l'attributo `@corresp`, il cui valore punta all'identificatore univoco della descrizione del corrispondente 'studio', codificata in `<genDesc>`.

Infine, come accennato nel paragrafo 3, una delle questioni di principale interesse nella codifica degli 'studi' riguarda la possibilità di realizzare un collegamento tra alcune informazioni contenute negli schizzi e altre presenti in fonti esterne, quali il consistente epistolario di Bellini. Si tratta, ad esempio, di richiami a opere e date, a passi specifici della musica belliniana, come arie o duetti precisi, o ancora a cenni che forniscono preziosi indizi sui processi compositivi adottati dal compositore. L'ipotesi che, allo stato attuale, si è ritenuta più promettente prevede la codifica di informazioni di questo tipo facendo riferimento al Web Annotation Data Model (WADM)[19]: l'esempio è fornito – ancora una volta – dal vocabolario TEI, che permette di creare il collegamento con un URI come target, corredato da metadati esplicativi sull'annotazione. Nell'esempio proposto in figura 3, si illustra una possibile modalità di collegamento tra l'epistolario nell'edizione BDC[20] [5] e la codifica del testo musicale, codificando l'informazione come nota critica in `<annot>`, un elemento del vocabolario MEI progettato sul modello di `<annotation>` in TEI. Il riferimento alla risorsa esterna è stato codificato come URI dell'attributo `@target` dell'elemento figlio `<ref>`, utilizzando un handle al repository dell'edizione BDC creato in CLARIN[21]. Gli attributi `@startid` e `@endid` permettono di contestualizzare l'annotazione con un luogo preciso della codifica degli 'studi' (nel caso dell'esempio proposto, i punti coincidono con il valore dell'identificatore univoco della prima e dell'ultima sillaba del testo lirico sottoposto allo 'studio' in questione).

```
<annot xml:id="link-BDC_1" startid="#m-347" endid="#m-393">
  <ref target="https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/bitstream/
  handle/20.500.11752/OPEN-1000/BDC-XML.zip?sequence=1/encoding-main/
  LL1_26.xml#LL1.26">Il libro ha il gran difetto che non è bene dialogato: le
  situazioni sono belle, l'espressioni ripetute, comuni, stupide qualche
  volta, in una parola si vede che chi l'ha scritto non avea ne cuore, ne
  cognizioni per bene esprimere i sentimenti dei suoi personaggi.</ref>
  <locus>LL1.26, carta 1r, righe 14 e segg.</locus>
</annot>
```

Figura 3. Estratto della codifica con riferimento a un luogo della lettera LL1.26 in [5]

L'ipotesi proposta è solo una delle tante possibili letture che possono essere adottate nell'ampio spettro offerto dal lavoro di marcatura del testo: si tratta di un risultato parziale, nell'auspicio di un costante arricchimento delle informazioni codificate, che porti con sé ulteriori riflessioni sulla rappresentazione del testo musicale nelle sue specificità, anche in funzione di una sua fruizione digitale[22].

BIBLIOGRAFIA

[1] Bellini, Vincenzo. *Carteggi*. (a cura di) Graziella Seminara. Historiae Musicae Cultores 131. Firenze: Olschki, 2017.

[2] Bellini, Vincenzo, e Carlo Pepoli. *I Puritani*. (a cura di) Fabrizio Della Seta. Vol. 10. Edizione critica delle opere di Vincenzo Bellini. Milano: Ricordi, 2013.

[3] Caraci Vela, Maria. *La Filologia Musicale. Istituzioni, storia, strumenti critici. Vol I: Fondamenti storici e metodologici della Filologia musicale*. Lucca: Libreria Musicale Italiana, 2005.

[4] Ciotti, Fabio. *Il testo e l'automa: saggi di teoria e critica computazionale dei testi letterari*. Roma: Aracne, 2007.

[5] Del Grosso, Angelo Mario, e Daria Spampinato, (a cura di). *Bellini Digital Correspondence*. CNR Edizioni, 2023.

[6] Mantica, Candida Billie. «Gli "studi giornalieri" di Bellini "sviluppati con effetto" nei Puritani». *Bollettino di Studi Belliniani* VI (2020): 29–73.

[7] Pierazzo, Elena, e Tiziana Mancinelli. *Che cosa è un'edizione scientifica digitale*. Roma: Carocci, 2020.

[18] «If interested in modeling such dependencies between witnesses, using markup from 3.5 *Functional Requirements for Bibliographic Records (FRBR)* is generally recommendable» (MEI Guidelines, *11.1.1 General Usage*, https://music-encoding.org/guidelines/v5/content/scholarlyediting.html#critAppElements).
[19] https://www.w3.org/TR/annotation-model/
[20] http://bellinicorrespondence.cnr.it
[21] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/OPEN-1000. Il riferimento è quindi un permalink sostenibile univoco e citabile globalmente.
[22] Un primo tentativo di visualizzazione dei dati codificati è stato realizzato tramite lo strumento MEI Viewer del tool open source Verovio (https://www.verovio.org/mei-viewer.xhtml).

[8] Roeder, Torsten, Fabian C. Moss, e Maik Köster. «Music-Text Interlinking as a Challenge for Digital Encodings of Music-Theoretical Writings». In *Encoding Cultures: Joint MEC TEI conference 2023 – Book of Abstracts*, (a cura di) Raffaele Viglianti, 172–73, 2023. https://doi.org/10.5281/zenodo.10427826.

[9] Seminara, Graziella, e Anna Tedesco, (a cura di). *Vincenzo Bellini nel secondo centenario della nascita. Atti del Convegno Internazionale*. Catania 8-11 novembre 2001: Olschki, 2001.

[10] Tancredi, Giulia, e Cristina Fenu. «XML-TEI: Un modello per la filologia d'autore». In *AIUCD 2022 - Culture digitali. Preceedings della 11a conferenza nazionale*, (a cura di) Fabio Ciracì, Giulia Miglietta, e Carola Gatto, 218–22. Lecce, 2022. https://conference.unisalento.it/ocs/index.php/aiucd2022/index/pages/view/proceedings.

[11] Viglianti, Raffaele, (a cura di). *Encoding Cultures: Joint MEC TEI conference 2023 – Book of Abstracts*, 2023. https://doi.org/10.5281/zenodo.10427826.

[12] Viglianti, Raffaele. «Music and Words: Reconciling Libretto and Score Editions in the Digital Medium». In *Ei, dem alten Herrn zoll' ich Achtung gern': Festschrift für Joachim Veit zum 60. Geburtstag*, (a cura di) Kristina Richts e Peter Stadler, 727–46. München: Allitera Verlag, 2016. https://doi.org/10.25366/2018.46.

Verismo digitale. Per un'edizione digitale commentata delle opere di Verga, Capuana, De Roberto

Liborio Pietro Barbarino[1], Elisa Conti[2], Christian D'Agata[3],
Miryam Grasso[4], Ninna Maria Lucia Martines[5], Eliana Vitale[6]
[1] Università di Catania, Italia - liborio.barbarino@unict.it
[2] Università di Catania, Italia - elisa.conti@phd.unict.it
[3] Università di Catania, Italia - christian.dagata@unict.it
[4] Università di Catania, Italia - miryam.grasso@unict.it
[5] Università di Catania, Italia - ninna.martines@phd.unict.it
[6] Università di Catania, Italia - eliana.vitale@unict.it

ABSTRACT
Il contributo propone una presentazione del progetto, finanziato nell'ambito del PE5 Changes, Spoke 3, *Verismo digitale*, finalizzato alla creazione di un ecosistema digitale intorno alle opere della letteratura verista. Verga, De Roberto, Capuana si mostrano attraverso edizioni digitali sinottiche, hyperedizioni, concordanze. Si discute di un modello d'uso del digitale che integri gli strumenti di filologia, lessicografia, ermeneutica e didattica, rivolgendosi tanto al lettore occasionale quanto allo studioso.

PAROLE CHIAVE
Digital Scholarly Edition; codifica XML/TEI; teoria e prassi del commento; lessicografia; ermeneutica.

1. INTRODUZIONE
Verismo digitale nasce nell'ambito del WP3 dello Spoke 3 del progetto PNRR CHANGES e all'interno del Dipartimento di Scienze Umanistiche dell'Università di Catania (responsabile scientifico dell'unità: Marina Paino, componenti: Andrea Manganaro, Antonio Sichera), dove rappresenta un'espansione e ulteriore raffinamento dei prodotti e delle competenze maturate all'interno del Cinum (Centro di Informatica Umanistica) in esperienze come (ma non esclusivamente) l'Edizione digitale dell'Opera Omnia di Luigi Pirandello[1] nel contesto dell'Edizione Nazionale delle opere dello scrittore. Obiettivo del progetto è quello di presentare edizioni scientifiche digitali [14] di testi rappresentativi della letteratura verista, secondo un concetto che vede il testo letterario come *hub* di una rete da cui si diramano varie esperienze di lettura, tra esse collegate: materiali preparatori, varianti d'autore, commenti ipertestuali, strumenti lessicografici, percorsi didattici. La collocazione del progetto è trasversale, poiché tali testi, filologicamente accurati e scientificamente trattati, potranno essere consultati dall'utente – lo studioso e il curioso, il docente e il discente universitario, o di scuola secondaria – in accordo con le sue esigenze. Il corpus di riferimento è dunque individuato nei tre maggiori autori del Verismo (Verga, Capuana, De Roberto), che saranno proposti – salvo motivate eccezioni – nelle *editiones principes* delle opere selezionate.
Le ragioni sottostanti a tale scelta sono molteplici: a) L'esperienza del Verismo è riconosciuta come essenziale nella storia letteraria italiana, per i suoi risultati sia in termini d'invenzione (*I Malavoglia* su tutti) che di riflessione metaletteraria: asse che congiunge la fondazione del genere romanzo in Italia, ad opera di Manzoni, con le esperienze che hanno segnato il Novecento (da Svevo a Pirandello), riguadagnando poi centralità nel dibattito e nella considerazione critica con le istanze realistiche emerse almeno a partire dal secondo dopoguerra. b) Se Manzoni, Svevo, Pirandello godono di accessibilità e appropriato trattamento[2] che gli permettono di raggiungere un pubblico trasversale, nulla di simile è stato finora realizzato per la letteratura verista[3]. c) La critica ha documentato il rapporto fra Verga, Capuana e De Roberto attraverso l'evidenza di una teoria di glosse, suggerimenti, revisioni e letture critiche che i tre hanno scambiato con costanza, giungendo a definirlo come un vero e proprio laboratorio di scrittura [10]. Tali esperienze letterarie potranno dunque venir fruite globalmente all'interno di un unico portale. d) A queste si aggiungano una considerazione di ordine pratico – si tratta di testi liberi dai diritti d'autore – e una di carattere glocale, che spinge una delle riconosciute istituzioni culturali del territorio a farsi promotrice e custode di una storia che proprio da queste latitudini prendeva le mosse.

[1] http://www.pirandellonazionale.it/.
[2] Per Manzoni si pensi a [https://www.alessandromanzoni.org/opere/1] e [https://projects.dharc.unibo.it/leggomanzoni/].
[3] Il VIVer (Vocabolario dell'Italiano Veristico), promosso dalla Fondazione Verga e dall'Accademia della Crusca, ha posto al centro della sua attenzione gli autori della letteratura verista per la costituzione di un «*corpus* dei *corpora*», in una prospettiva squisitamente lessicografica. Tra gli sviluppi futuri andrà certamente considerato di far riferimento alle acquisizioni del vocabolario come database esterno per i testi di *Verismo digitale*.

2. METODI

La realizzazione di un'edizione digitale commentata presuppone un approccio storico-letterario e critico-ermeneutico che consenta di ancorare tale pratica a presupposti scientifici attendibili. In questa prospettiva, il commento diventa fondamento di ogni atto critico-interpretativo e ponte di un contatto profondo, scevro da condizionamenti, fra il testo e il lettore, utile a «costruire un'educazione alla fruizione critica» [6: 246].

Se è al lettore e al suo tempo che bisogna guardare nella restituzione della potenzialità semantica di un testo, il progetto deve prendere atto delle conseguenze della transizione digitale sull'evoluzione storica e sociologica della pratica della lettura, che garantisce una fruibilità libera e in *open access*, ma spesso incide sulla scientificità e affidabilità delle risorse (disponibili a lettori non sempre in possesso di strumenti critici adeguati a un approccio autonomo ai testi e alle fonti).

Nella prassi del commento ai testi letterari, il paradigma digitale del progetto, collocandosi tra rispetto fedele della tradizione e apertura all'innovazione, si configura allora come risorsa atta a favorire la ricezione testuale per un pubblico di lettori sempre più ampio e non selezionato *a priori*, al quale sia data la possibilità di interrogare i testi veristi e di ricavarne delle risposte individuali, contando sull'autorevolezza di un *background* critico scientificamente fondato e sulla trasparenza nella presentazione dei dati. Solo così sarà possibile garantire una fruizione del testo letterario consapevole, non distorta e al contempo autonoma.

In concreto – e secondo uno schema ormai consolidato per la ricerca umanistica in ambito digitale –, la modellizzazione dei testi degli autori prescelti e del commento si è servita dello schema di codifica proposto dalla *Text Encoding Initiative* (TEI). Tale scelta è finalizzata a favorire il riuso, possibili future implementazioni, la non obsolescenza e l'interoperabilità dei dati in accordo con i principi FAIR[4].

I testi saranno presentati secondo *facies* differenti – edizione facsimile; edizione sinottica di differenti redazioni di un testo; hyperedizione con commento; concordanza –, curvate secondo le peculiarità di ciascun testo e autore, e visualizzate con un'interfaccia ad hoc sviluppata all'interno del progetto. Verranno contestualmente utilizzati dei software *open access*, quali EVT[5], per alcune specifiche edizioni scientifiche digitali.

Nello specifico del commento, un primo livello – denotativo – è stato prodotto attraverso dei tag di tipo strutturale (`<div>`, `<p>`, e `<pb>`) e la marcatura delle entità nominate (`<persName>`, `<placeName>`, `<orgName>`, `<objectName>`, `<rs>`). Tale *step* – ormai pacificamente considerato punto di partenza per le successive analisi ed elaborazioni [15] – è arricchito con il rilievo di elementi come il discorso diretto o le date, insieme ad altri esemplati sul particolare testo o autore (discussi *infra* nel dettaglio). Si è poi aggiunto un secondo livello concernente in modo più specifico la comprensione del testo, che accoglie delle note di approfondimento di carattere filologico, lessicografico, metacritico. Tale secondo livello di commento è stato realizzato allestendo dei pointer (`<ptr>`) – con `@type` e `@subtype` specifici – al *back* del documento, nel quale sono presentate le corrispondenti note filologiche, lessicografiche e bibliografiche. Su alcuni passi notevoli, è previsto un piano ulteriore, che – portando a sintesi la stratificazione dei primi due – arrivi a illuminare per tratti condensati ma profondi le verticali semantiche dell'opera. Se i primi due livelli riguardano dunque l'aspetto della spiegazione puntuale, del dato anche bibliografico, è quest'ultimo che mira invece a una più ampia dimensione ermeneutica.

3. ARTICOLAZIONE DEL PROGETTO

Si entra adesso nel merito di alcune scelte e delle rispettive implicazioni riguardanti i lavori condotti sui singoli autori: Giovanni Verga (*I Malavoglia*, le *Novelle rusticane*), Luigi Capuana (*Le fiabe*, *I racconti*), Federico De Roberto (*I Viceré*).

3.1. Giovanni Verga

3.1.1. L'hyperedizione dei *Malavoglia*: un prototipo

Il testo della *princeps* (Treves 1881) – emendato dai refusi e dalle sviste, secondo il criterio di cautela espresso e documentato dall'autorevole edizione critica dell'opera [20: LXIII-XC] – costituisce insieme il *corpus* da lemmatizzare al fine di produrre una concordanza dell'opera, e il *body* da sottoporre a codifica. La concordanza rappresenta sia un avanzamento di per sé (non esistono concordanze dell'opera), che il tessuto da cui possono ricavarsi i dati lessicografici che contribuiranno a strutturare il commento. La codifica come detto, muove invece da una base strutturale, integrata con la marcatura delle *Named Entities* e di altri elementi pensati e discussi con il gruppo di lavoro proprio a partire dalle questioni poste dal testo verghiano: lessico, discorso diretto, date. Vediamo adesso nel dettaglio come tali scelte implichino

[4] Fair Principles: https://www.go-fair.org/fair-principles/
[5] EVT: http://evt.labcd.unipi.it/

già delle attese di carattere ermeneutico, e si connettano dunque al livello più profondo della ricerca, quello del commento nell'hyperedizione.
La marcatura di elementi distinti del lessico verghiano (`<distinct>`) risulta funzionale sia in vista di una puntuale *explicatio* (cosa significa «alare una parommella»? cosa sono i «lupini»?) che di un livello ulteriore nel quale lo specialista può trovare utili riscontri bibliografici (l'enorme mole di studi sull'opera invita a questo), prima di convergere trasversalmente su quanto dirimere il punto in questione possa incidere sulla comprensione profonda dell'opera (perché i lupini non possono essere frutti di mare?).
Segnalare i passaggi del discorso diretto – utilizzando il tag `<q>` insieme agli attributi `@who` e `@toWhom` – consente di separare quanto certamente attribuibile a un personaggio (perché segnalato tramite interpunzione) da altri elementi che con diversi gradienti ricadono nell'ambito del narratore. Si restringe così il campo intorno alla proteiforme figura del narratore verghiano (con l'intento di migliorarne la definizione); e insieme si raccolgono dati (una sorta di concordanza selettiva) che possono essere visualizzati per meglio comprendere (e rappresentare) le relazioni tra i personaggi, il lessico specifico di ciascuno (esiste? è diverso da quello degli altri o del narratore?).
Un'indagine stringente sui passaggi cronologici dell'opera (tag `<date>` con attributo `@when`) consente di mettere a fuoco una puntuale cronologia degli eventi e di vagliarla sia nel merito che in relazione agli appunti preparatori dell'autore. Nel capitolo iniziale il lettore si imbatte in una data certa (il dicembre del 1863 in cui 'Ntoni è chiamato per la leva di mare, che costituisce l'antefatto delle vicende narrate) e poi in due eventi-chiave di collocazione più sfumata: il «negozio dei lupini» e la successiva partenza della *Provvidenza* (la barca di famiglia che naufragherà insieme al carico e a Bastiano Malavoglia). Secondo la tradizione, il primo sarebbe da collocare domenica 8 settembre [23: 24] mentre la seconda, attraverso gli appunti dell'autore [22: 347], sabato 21 settembre. Nei calendari storici, tuttavia, il primo anno successivo al 1863 in cui a una domenica 8 settembre segua un sabato 21 settembre, sarebbe un 1867 del tutto incongruo con le vicende narrate[6]. Anche in forza della rappresentazione di questi dati, il commento digitale può agevolmente evidenziare la consistenza simbolica della temporalità in un testo ritenuto (peraltro giustamente) un capolavoro della letteratura realistica.

3.1.2. Per un'edizione digitale commentata delle *Novelle rusticane*

La raccolta delle *Novelle rusticane*, pubblicata integralmente per la prima volta dall'editore Casanova di Torino nel dicembre del 1882, con la data del 1883 [21: LI], rappresenta un'opera notevole nella produzione verghiana della stagione verista che riconduce a sintesi i nodi concettuali e di poetica della produzione precedente per transitare verso la stagione successiva, dalle ultime *correspondances* simboliche dei *Malavoglia* [13: 172] allo spietato materialismo borghese del *Mastro-don Gesualdo*. Come aveva già notato Luigi Russo nella sua *Prefazione* all'edizione fiorentina Vallecchi del 1924, la raccolta costituisce infatti un significativo passaggio logico, e non solo cronologico, tra i due romanzi maggiori di Verga [19: VII].
Tenuto conto della tradizione interpretativa e, ovviamente, dell'accuratissima edizione critica delle *Novelle rusticane* a cura di Giorgio Forni (pubblicata nel 2016 per l'Edizione nazionale delle Opere di Giovanni Verga, diretta da Gabriella Alfieri) [21], il progetto di un'edizione digitale commentata e a misura di lettore si rivela ancora più utile e funzionale nel caso di una raccolta come le *Novelle rusticane*, complessivamente non ancora pienamente valorizzata nell'alveo della produzione verghiana e spesso ridotta a uso scolastico in antologie che propongono alle lettura soltanto alcuni dei racconti che la compongono (*La roba* e *Libertà* tra le letture più comuni).
Si procede in primo luogo con la codifica in XML/TEI dell'*editio princeps* delle *Novelle rusticane* (1883), ma si registrano anche le varianti significative ai fini del commento – ovvero quelle rilevanti sul piano narratologico – delle novelle dalle prime redazioni per la pubblicazione su rivista all'edizione «definitiva riveduta e corretta dall'autore», pubblicata per «La Voce» di Prezzolini nel 1920 [20].
La «critica digitale delle varianti», ossia «l'uso di metodi di indagine computazionale per l'analisi delle varianti» [12: 42], è finalizzata a interpretare le modifiche apportate al dato testuale, in sincronia e in diacronia, come spie dell'evoluzione delle strategie narratologiche di Verga in relazione a fattori intrinseci alla maturazione della poetica autoriale, nonché di natura editoriale o storico-culturale.
La codifica avviene sulla base dei presupposti condivisi dalle unità del progetto con un grado di personalizzazione motivata, come annunciato, dalle esigenze interpretative dettate dal testo. La fase della marcatura permette infatti di ingaggiare un corpo a corpo con il testo dal quale emergono questioni cruciali nella direzione di un commento metatestuale. Le due attività critiche sono dunque interconnesse e funzionali l'una all'altra. A tal proposito, si fornisce un esempio applicativo

[6] Il punto in questione è dibattuto in relazione al cap. IX, ma non al primo, e riguarda l'impossibile coincidenza tra la festa per il fidanzamento di Mena (in giugno) e la notizia della morte di Luca nella battaglia di Lissa, storicamente collocata il 20 luglio del 1866.

della metodologia adottata, riportando la codifica di una porzione testuale della novella *La roba*, sintatticamente e semanticamente rilevante.

```
Ed anche la <distinct type="keyword">roba</distinct> era fatta per lui, che pareva
ci avesse la calamita, perché <seg type="proverbio">la roba vuol stare con chi sa
tenerla</seg><ptr type="compr" subtype="fraseologia" target="N_RU_FR_roba"/>, e
non la sciupa come quel barone che prima era stato il padrone di <persName
ref="#Mazzarò">Mazzarò</persName>
```

Il commentatore, forte della conoscenza critica del *corpus* verghiano, oltre a marcare il nome del protagonista della novella, Mazzarò (tag `<persName>` e attributo `@ref="#Mazzarò"` per l'identificativo del personaggio), non può non individuare nel caso proposto l'occorrenza di una parola chiave del lessico verghiano, la "roba" (tag `<distinct>` e attributo `@type="keyword"`) e di una componente fraseologica rilevante, il proverbio «la roba vuol stare con chi sa tenerla». Nella codifica il proverbio viene marcato associando a esso un *pointer* `<ptr>` che rimanda a una lista di proverbi fornita nel *back*. La codifica così strutturata, integrata all'allestimento e allo studio delle concordanze, permette di ricavare dal *corpus* – con riferimento al caso specifico – repertori di personaggi, di elementi lessicali significativi e di proverbi, a partire dai quali è possibile proporre interpretazioni ancorate alla dimensione testuale, ma proiettate a quella contestuale.
Il progetto prevede inoltre la riproduzione anastatica dei disegni di Alfredo Montalti, apparato iconografico dell'*editio princeps* delle *Novelle rusticane* [24], per riflettere su forme e funzioni delle illustrazioni nel contesto del racconto verista.

3.2. Luigi Capuana

3.2.1. *Le fiabe* di Luigi Capuana

Le raccolte *C'era una volta*, *Il Raccontafiabe* e *Chi vuol fiabe, Chi vuole?* sono i testi fiabistici di Capuana selezionati per il portale *Verismo digitale*. Rispetto agli altri testi presentati, le fiabe del mineolo risultano certamente peculiari, rappresentative di un genere che sembra divergere rispetto alla 'letteratura del Vero'. Nonostante la natura intrinseca di questa tipologia di opere, Capuana mette a tema l'impossibilità di un cambiamento nello *status quo*: solo chi è predestinato può diventare un Reuccio, chi invece non lo è resterà aggiogato all'umiltà della sua condizione [2]. In tal senso, dunque, la presenza delle fiabe risulta pienamente in contesto.
Proponiamo di seguito un breve affondo volto a chiarire quali sono gli strumenti impiegati per questo oggetto di analisi così particolare e quali le strategie adottate nell'ambito della codifica. Facciamo riferimento al lavoro svolto e progettato per *C'era una volta*, la prima raccolta di fiabe di Capuana, pubblicata nel 1882 con dodici testi, che ebbe diverse riedizioni, fino all'ultima del 1889 che consta di sedici fiabe. È questa, per la sua maggiore completezza, la lezione messa a testo per la nostra edizione, contrariamente a quanto deciso in relazione agli altri testi. Tutte le fiabe sono state oggetto di un processo correttorio da parte dell'autore, che è intervenuto sia in ambito linguistico che nella strutturazione dei testi e del macrotesto. Capuana risistema le sue fiabe, attingendo dal corpus iniziale del 1882 e aggiungendone dalla raccolta *Il Regno delle Fate* (1883). Due questioni fondamentali per analizzare l'evoluzione delle fiabe sono dunque i cambiamenti nella struttura delle raccolte e l'evoluzione del linguaggio. Nel passaggio dalla *princeps* all'ultima edizione l'autore ricerca una lingua letteraria indirizzata al nuovo popolo italiano, esemplata su un lessico toscaneggiante vicino a quello adoperato da Collodi [9].
Secondo quest'ottica, riteniamo rilevante uno studio delle varianti del testo – messe a fuoco nell'edizione critica digitale, codificata in XML/TEI e con visualizzazione in EVT2 –, anche nell'ottica di un'analisi dell'evoluzione del lessico, indagata attraverso l'hyperedizione con commento e lo strumento delle concordanze.
Per quanto riguarda l'edizione filologica, dopo aver studiato le caratteristiche delle fiabe, è stato necessario ottenere i testi in formato elettronico dalle edizioni scelte e collazionare i passi. Il software Abbyy FineReader[7] è stato utilizzato per il riconoscimento ottico dei caratteri. Il passo successivo è stato il confronto dei testi, tramite l'impiego del software di collazione automatica LERA[8]. Si tratta di un software online, realizzato dall'Università di Halle-Wittenberg [15], che offre la possibilità di collazionare più testi, caricando i file delle varie edizioni in formato .txt o .xml. Una volta effettuato l'upload dei testi, è poi possibile scegliere la modalità di distinzione delle varianti e soprattutto il formato con cui scaricare l'output [18]. Marcus Pöckelmann e il suo team, tramite il loro software, offrono la possibilità di ottenere il testo in codifica XML/TEI, programmato per TEI Publisher. Nel nostro caso, il tipo di codifica è differente, ma è bastata una semplice revisione per ottenere un file codificato secondo la modellizzazione fatta.
All'interno dell'edizione filologica si propongono, oltre alla lezione a testo del 1889 (C89), tre testimoni: *C'era una volta* del 1882 e del 1885 (rispettivamente C82 e C85) e *Il regno delle fate* del 1883 con sigla RFT83.

[7] Abbyy FineReader: https://pdf.abbyy.com/it/
[8] LERA software: https://lera.uzi.uni-halle.de/

Il set di marcatori utilizzati appartiene al modulo TEI *textcritic* e sono <app>, <lem> e <rdg>:

```
<app>
    <lem wit="#C89 #C85">gallettina</lem>
    <rdg wit="#C82">gallinetta</rdg>
</app>
```

Si tratta di un set di marcatori che può essere impiegato nel caso di varianti circoscritte, ma non per descrivere trasposizioni di interi blocchi di testo. Poiché Capuana interviene massicciamente nell'impianto delle raccolte, attingendo testo da una fiaba e spostandolo in un'altra, abbiamo scelto di indicare tale casistica attraverso il tag <note>. Non si tratta dunque di una marcatura ermeneutica, ma di un espediente per segnalare la questione al lettore.
Per quanto riguarda il secondo tipo di edizione (Hyper), la codifica delle Entità Nominate – in linea con le altre opere proposte all'interno del portale – è stata modellata al fine di dare visibilità a particolari tipi di oggetti (<objectName>) come quelli magici. Ai luoghi è stato assegnato il tag <placeName>, precisando che si tratta di luoghi di fantasia, non reali, e senza una specifica denominazione. Ad esempio, tra le fiabe del mineolo non si trova mai il luogo del castello, bensì sempre del palazzo reale [5] cosa che produce un chiaro scostamento dall'immaginario medievale prevalente nel genere, e dunque può configurarsi come un interessante oggetto d'indagine.

3.2.2. *I racconti* di Luigi Capuana

In parallelo con la scrittura fiabesca, il percorso creativo di Capuana «si definisce anzitutto sul piano della prassi novellistica» [17: 20]. Per tale motivo, fra i testi selezionati per il portale si è scelto di includere le seguenti raccolte, ciascuna rappresentativa di uno dei tre filoni della produzione novellistica capuaniana: *Le Paesane* (ambientazione rusticana), *Le Appassionate* (indagine psicologica) e *Storia fosca* (che testimoniano l'interesse dell'autore per l'occultismo). Concordemente a quanto già espresso sono previsti due livelli di codifica: oltre al 'commento' che prevede la marcatura delle *Named Entities*, risulta particolarmente rilevante e necessaria la codifica di tipo filologico, poiché non esiste ancora un'edizione critica dei racconti di Capuana. Le vicende editoriali vissute da queste raccolte sono particolarmente complesse: dopo una prima pubblicazione su rivista, le novelle furono ripubblicate all'interno di diverse edizioni in volume, con l'introduzione di varianti nel passaggio da un'edizione all'altra.
A seguito di un'indagine preliminare sui testimoni dei racconti manoscritti e a stampa, condotta anche presso l'Archivio Capuana della Biblioteca comunale di Mineo, si è scelto di includere nell'edizione solo i testimoni a stampa. Sono in effetti questi, più della tradizione manoscritta, a dare conto dell'evoluzione linguistica e strutturale dei testi.
Il lavoro finora svolto ha riguardato le novelle delle *Paesane*. I testimoni collazionati sono contenuti nella sezione <SourceDesc> del *Tei Header*, all'interno dell'elemento <listWit>. Per l'apparato, come per le fiabe, si adopera il *Parallel Segmentation Method*[9], con l'esplicitazione di ciascuna variante in parallelo.
I problemi maggiori nell'utilizzo di questo modello di codifica sono stati posti da *Nostra gente* e *Dalla terra natale*, edizioni tardive nelle quali l'autore non interviene minutamente su punteggiatura, singole parole o sintagmi, ma su interi blocchi, alcuni dei quali subiscono una riscrittura pressoché integrale[10]. L'ipotesi di trascrivere per intero la versione riportata da ogni testimone si scontra con il fatto che il testo delle edizioni precedenti è molto vicino a quello delle *Paesane*. Dunque, codificare le varianti riprendendo di volta in volta l'intera porzione di testo renderebbe il codice ridondante, oltre a comportare importanti ricadute sulla consultazione dell'edizione. Il software di visualizzazione evidenzierebbe infatti intere porzioni di testo, aspetto che renderebbe molto difficoltoso per il lettore individuare a colpo d'occhio le varianti che sono invece presenti a livello microtestuale.
Si è scelto pertanto di adottare una soluzione di questo tipo:

[9] TEI Consortium, Guidelines 12.2.3: https ://tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS.
[10] Prendiamo ad esempio l'*incipit* della novella *Lo sciancato*, particolarmente rappresentativa di questi aspetti. Nell'edizione delle *Paesane* leggiamo: «Da bimbo, nel saltare un muricciolo, s'era rotta una gamba, e il dottore gliel'aveva rimessa così male che gli era rimasta quasi due dita più corta dell'altra. Dal giorno che l'avevano visto arrancare un po' contorto dal lato destro, non l'avevano più chiamato col suo nome; e, dopo, se uno avesse domandato di Neli Frisinga, tutti gli avrebbero risposto che non lo conoscevano e non l'avevano neppur sentito nominare in Mineo [3]. La versione di Nostra gente recita invece: «Lo avevano chiamato sin da ragazzo lo Storto perché era nato con una gamba più corta dell'altra; e dopo, se uno avesse domandato di Neli Frisinga, tutti gli avrebbero risposto: – Non lo abbiamo mai sentito nominare in Mineo». [4].

```
<app>
    <lem>
        <app>
            <lem wit="#P #H1 #H2">Da bimbo</lem>
            <rdg wit="#Riv">Bimbo</rdg>
        </app>
        , nel saltare un muricciolo, s'era
        <app>
            <lem wit="#P">rotta</lem>
            <rdg wit="#Riv #H1 #H2">rotto</rdg>
        </app>
        una gamba, e il dottore gliel'
        <app>
            <lem wit="#P #Riv #H1">aveva</lem>
            <rdg wit="#H2">avea</rdg>
        </app>
        <app>
            <lem wit="#P #H1 #H2">rimessa</lem>
            <rdg wit="Riv">rappiccata</rdg>
        </app>
        così male che gli era rimasta
        <app>
            <lem wit="#P #H1 #H2">quasi due dita più corta</lem>
            <rdg wit="#Riv">più corta quasi due dita</rdg>
        </app>
        dell'altra.
    </lem>
      <rdg wit="#NG">Lo avevano chiamato sin da ragazzo lo Storto perché
era nato con una gamba più corta dell'altra;</rdg>
</app>
```

Si indica con <rdg> la versione di #NG, molto diversa rispetto a quella degli altri testimoni. All'interno dell'elemento <app> viene quindi annidato un secondo elemento <app> che contiene le varianti dei gruppi dei testimoni più simili tra loro, individuando con <lem> la versione del testo base #P e dei testimoni che con essa concordano, e con <rdg> le varianti dei testimoni che invece se ne discostano per leggere modifiche. Oltre a rendere il codice meno prolisso, si ha un vantaggio nella rappresentazione: è sufficiente modificare il foglio di stile perché i due livelli di apparato vengano evidenziati con colori diversi, rendendo visibili anche le varianti che interessano la punteggiatura o le singole parole.

3.3. *I Viceré* di Federico de Roberto

Sulla scia di Giovanni Verga e di Luigi Capuana, Federico De Roberto si inserisce nell'alveo dell'esperienza verista mostrando un dettato perfettamente in grado di distinguersi da quello dei 'maestri'. *Verismo digitale* prevede la codifica del più celebre dei romanzi derobertiani, *I Viceré*. Esso presenta due edizioni: l'*editio princeps* del 1894, pubblicata dall'editore Galli di Milano [7], e l'edizione Treves [8] pubblicata, sempre a Milano, nel 1920. Il testo di base segue, secondo il criterio sopramenzionato, quello della *princeps* del 1894.

Così come per Verga e Capuana, anche per De Roberto l'obiettivo è stato elaborare un ecosistema digitale dalla visualizzazione 'multiprospettica' costruito su tre livelli: l'hyperedizione, le concordanze e un prototipo di visualizzazione dinamica in EVT2 delle varianti a stampa dell'edizione Galli del 1894 e dell'edizione Treves del 1920. Si condividono pertanto i primi avanzamenti del progetto, specificando quali peculiarità del testo sono state messe in evidenza nel commento.

Dopo la fase di riconoscimento del testo (OCR) a partire dalle immagini ad alta definizione della princeps – e successivamente ai diversi cicli di ricontrollo –, è stata avviata la codifica con l'obiettivo di rappresentare i diversi livelli (strutturale, entità nominate, semantico) allestendo l'edizione scientifica digitale commentata.

Si precisa che la codifica non ha l'obiettivo di ricostruire a tutto tondo il processo scrittorio del romanzo, ma intende piuttosto dare notizia dei fenomeni variantistici più significativi. Per un'idea del travagliato iter compositivo dell'opera si rimanda ad Amaduri [1] e Morace [16].
Un peso significativo hanno piuttosto le note di approfondimento storico, specie quelle relative alla questione risorgimentale in Sicilia e al conseguente trasformismo politico, tematiche portanti del romanzo.
I tag utilizzati sono in particolare <date> per annotare le date, <event> per rappresentare gli eventi storici fondamentali e dei *pointer ad hoc* – per approfondire alcuni segmenti testuali con delle note di carattere storico:

```
<p>"Chi?… Quando?… La <placeName ref="#Francia">Francia</placeName>? Bel servizio! Bell'aiuto!… <persName ref="#Garibaldi">Garibaldi</persName>? Chi è <persName ref="#Garibaldi">Garibaldi</persName>? Non lo conosco!…»</p>
<p>Imparò a conoscerlo il <date when="1860-05-13" ana="#VI_EV_Sbarco">13 maggio</date>, quando scoppiò come una bomba la notizia dello sbarco di <placeName ref="#Marsala">Marsala</placeName>.<ptr type="compr" subtype="storia" target="#N_VI_ST_SbarcoMarsala"/></p>
```

Come si nota dall'esempio sopra riportato, eventi come lo sbarco di Marsala del 13 maggio 1860 vengono codificati su due livelli: da un lato, quando presente, con il riferimento puntuale alla singola data di cui viene esplicitato l'evento correlato (con il tag <date>, il cui attributo si riferisce al tag <event> presente nel *back*), dall'altro lato con una nota di approfondimento storico più distesa formulata con il <ptr> che rimanda alla nota corrispondente.
Inoltre, mediante il tag <media> è stato messo a punto un apparato iconografico in corrispondenza dei luoghi del testo che meglio si prestano ad essere accompagnati da un supporto di tipo multimediale (nomi di luoghi e di monumenti, nomi di personaggi storici, riferimenti a usi e costumi locali, riferimenti ad eventi storici dalla forte iconicità, ecc.). Ad esempio, nell'architettura paesaggistica del romanzo, un luogo dalla forte connotazione simbolica e soprattutto dalla complicata valenza politica, è il Monastero dei Benedettini di Catania. I tag 'multimediali' hanno consentito pertanto di restituire al testo una terza dimensione, dando immediata concretezza visiva alle potenti descrizioni derobertiane.
Quest'ultimo aspetto rientra dunque nell'obiettivo precipuo del progetto, cioè quello di fornire una visione sinottica e al contempo analitica del testo, a seconda della tipologia di interrogazione avviata dall'utente, il tutto in una dimensione che si vuole sempre in dialogo con la didattica.

4. IMPATTO ATTESO

Edizioni scientifiche, *corpora* digitali, strumenti lessicografici concepiti in relazione tra loro grazie a una codifica su più livelli daranno al lettore la possibilità di mettere in questione e visualizzare le relazioni semantiche dentro e fra i testi, di farne materiale didattico e di ricerca, di curiosità o di approfondimento.
Puntando a una profonda integrazione fra le risorse in un ambiente di fruizione/consultazione amichevole, il portale intende posizionarsi come nodo di un più ampio ecosistema digitale volto a preservare e promuovere il patrimonio della nostra letteratura, modello di una creazione di saperi virtuosamente espandibile ad altri *asset* della cultura.

BIBLIOGRAFIA

[1] Amaduri, Agnese. *L'Officina de I Viceré. La genesi del romanzo attraverso l'epistolario di Federico De Roberto*. Soveria Mannelli: Rubbettino, 2017.

[2] Barsotti, Anna. «C'era una volta...» il Verismo. Sulla fiabistica di Luigi Capuana». In *Capuana verista: atti dell'incontro di studio, Catania: 29-30 ottobre 1982*, 85–99. Catania: Fondazione Verga, 1982.

[3] Capuana, Luigi. *Le Paesane*. Catania: Giannotta, 1894.

[4] Capuana, Luigi. *Nostra gente*. Milano-Palermo-Napoli-Genova: Remo Sandron, 1915.

[5] Capuana, Luigi. *Stretta la foglia, larga la via. Tutte le fiabe*. (a cura di) Rosaria Sardo, con illustrazioni di Lucia Scuderi. Roma: Donzelli Editore, 2015.

[6] Cataldi, Pietro. «Commento e parafrasi». In *La strana pietà. Schede sulla letteratura e la scuola*, (a cura di) Pietro Cataldi, 233–48. Palermo: Palumbo, 1999.

[7] De Roberto, Federico. *I Viceré*. Milano: Galli, 1894.

[8] De Roberto, Federico. *I Viceré*. Milano: Treves, 1920.

[9] Fedi, Roberto. «Capuana scrittore di fiabe e la formazione di C'era una volta...» In *L'illusione della realtà: studi su Luigi Capuana*, (a cura di) Michelangelo Picone e Enrica Rossetti, 205–20. Roma: Salerno Editrice, 1990.

[10] Giuffrida, Milena. «Elementi di antropologia religiosa nell'opera di Capuana». In *Letteratura e antropologia. Generi, forme e immaginari. Atti del XXI Convegno Internazionale della MOD 13-15 giugno 2019*, (a cura di) Alberto Carli, Silvia Cavalli, e Davide Savio, 131–38. Pisa: Edizioni ETS, 2021.

[11] Italia, Paola. «Editing 2.0. Quali testi leggiamo e leggeremo in rete?» *Nuovi Argomenti*, marzo 2016.

[12] Italia, Paola. «Per una critica delle varianti digitale». (a cura di) Margherita De Blasi, 41–58. Napoli: Unior Press, 2023.

[13] Luperini, Romano. «L'allegoria di Gesualdo». In *Giovanni Verga. Saggi (1967-2018)*, (a cura di) Romano Luperini, 169–87. Roma: Carocci, 2019.

[14] Mancinelli, Tiziana, e Elena Pierazzo. *Che cosa è un'edizione scientifica digitale*. Roma: Carocci, 2020.

[15] Montemagni, Simonetta. «Trattamento automatico del linguaggio e Digital Humanities: metodi e strumenti, sfide». In *Digital Humanities. Metodi, strumenti, saperi*, (a cura di) Fabio Ciotti, 164–81. Roma: Carocci, 2023.

[16] Morace, Aldo Maria. «'Protostoria' dei Viceré». *Studi e problemi di critica testuale* 101, fasc. Fasc. 2 (2020): 67–113.

[17] Muoio, Ilaria. *Capuana e il modernismo*. Pisa: Pacini, 2023.

[18] Pöckelmann, Marcus, André Medek, Jörg Ritter, e Paul Molitor. «LERA—an interactive platform for synoptical representations of multiple text witnesses». *Digital Scholarship in the Humanities* 38, fasc. 1 (giugno 2022): 330–46. https://doi.org/10.1093/llc/fqac021.

[19] Russo, Luigi. «Prefazione». In *Novelle rusticane*, di Giovanni Verga, V–XXX, Firenze: Vallecchi, 1924.

[20] Verga, Giovanni. *I Malavoglia*. (a cura di) Ferruccio Cecco. Novara: Interlinea, 2014.

[21] Verga, Giovanni. *I Malavoglia. Testo critico e commento*. (a cura di) Ferruccio Cecco. Torino: Einaudi, 2014.

[22] Verga, Giovanni. *Novelle rusticane*. Torino: Casanova, 1883.

[23] Verga, Giovanni. *Novelle rusticane*. Roma: La Voce, 1920.

[24] Verga, Giovanni. *Novelle rusticane*. (a cura di) Giorgio Forni, con disegni di Alfredo Montalti. Edizione Nazionale delle Opere di Giovanni Verga, n.S. III. Novara: Fondazione Verga - Interlinea, 2016.

Verso l'Hyperedizione. Lo sviluppo di Pirandello Nazionale tra didattica e ricerca

Milena Giuffrida[1], Christian D'Agata[2], Giulia Cacciatore[3], Fabrizio Lo Presti[4]
[1] Università di Catania, Italia - milena.giuffrida@unict.it
[2] Università di Catania, Italia - christian.dagata@unict.it
[3] Università di Catania, Italia - giuliacacciatore83@gmail.com
[4] Università di Catania, Italia - fabrizio.lopresti1998@gmail.com

ABSTRACT
Pirandello Nazionale è il portale che dal 2017 ospita l'Edizione digitale dell'Opera Omnia di Luigi Pirandello. Ideato come un *knowledge site*, si propone di rispondere alle esigenze delle diverse tipologie di lettori di Pirandello, perseguendo l'attuazione di un criterio di usabilità scalabile in base al target d'utenza. Una sezione del portale di recente sviluppo è quella dell'Hyperedizione, realizzata in maniera sperimentale su *Il fu Mattia Pascal* e sull'*Enrico IV,* ma che funge da modello per l'implementazione del progetto. Se ne descrivono qui le caratteristiche e le ricadute, sia sul piano della ricerca che su quello della didattica. A quest'ultimo aspetto viene dedicato un focus su alcune esperienze che hanno visto gli studenti protagonisti attivi nella stesura del commento digitale.

PAROLE CHIAVE
Edizioni scientifiche digitali; Filologia e lessicografia; Luigi Pirandello; Didattica digitale; Editoria ed archivi digitali.

1. INTRODUZIONE
Pirandello Nazionale (https://www.pirandellonazionale.it/) è il portale che ospita l'Edizione digitale dell'Opera Omnia di Luigi Pirandello. Nato nel 2017 come zona complementare all'Edizione nazionale[1], è stato ideato da un'*équipe* del Centro d'Informatica Umanistica (CINUM) dell'Università di Catania, guidata da Antonio Sichera e Antonio Di Silvestro. Sin dalla nascita, il portale è stato progettato seguendo un approccio *reader oriented* [5], volto quindi a rispondere alle esigenze delle diverse tipologie di lettori di Pirandello (studiosi, docenti/studenti, appassionati), perseguendo l'attuazione di un criterio di usabilità scalabile in base al target d'utenza [3]. Proprio per questo *Pirandello Nazionale* si è discostato da subito dalla seppur giovane tradizione delle DSE, andando oltre la sola rappresentazione documentale in favore dell'interpretazione testuale. Questo tipo di approccio, che ha reso *Pirandello Nazionale* a tutti gli effetti un *knowledge site* [3 e 5], non lo ha però relegato al ruolo di semplice 'espansione digitale' del prodotto cartaceo (a sua volta caratterizzato dalla scalabilità d'uso, determinata dalla compresenza di un'*editio maior* – tradizionale edizione critica destinata a studiosi e biblioteche – e di un'edizione *prêt-à-porter*, quella degli Oscar Mondadori [8]), ma anzi ne ha fatto un ecosistema digitale «la cui natura e il cui scopo si collocano al crocevia tra portale multidisciplinare, edizione e archivio» [3]. Se questo approccio, insieme all'accessibilità (ogni contenuto è proposto in modalità Open Access, facilmente scaricabile in formato PDF) e alla scientificità dei contenuti sono stati da subito i punti di forza di *Pirandello Nazionale*, è a partire dal 2020 che si è cercato di superare le criticità in termini di condivisione, diffusione e replicabilità dei risultati effettivi della ricerca, al fine non solo di un adeguamento ai principi FAIR [11], ma anche di una maggiore interazione tra tutte le sezioni del portale. Il frutto di questa riflessione è oggi visibile nelle Hyperedizioni de *Il fu Mattia Pascal* e dell'*Enrico IV*, sezioni sperimentali dell'edizione digitale e modello per l'implementazione del progetto. La costituzione dell'Hyperedizione è stata determinante anche in funzione delle ricadute didattiche del portale, non solo per quanto riguarda la fruizione dei contenuti ospitati, ma soprattutto per la possibilità di avvicinare gli studenti ai linguaggi informatici, la conoscenza dei quali – oltre a essere pressoché indispensabile – comporta un indubbio vantaggio al fine dell'analisi del testo (come ben dimostrano i casi illustrati in [2]).

[1] La Commissione per l'Edizione Nazionale dell'Opera Omnia di Luigi Pirandello è stata nominata dal MIBACT nel 2017 ed è composta da studiosi italiani e stranieri, specialisti dell'opera di Pirandello; li elenchiamo di seguito: Angelo Pupino (Università di Napoli - Presidente), Aldo Maria Morace (Università di Sassari - Segretario), Beatrice Alfonzetti (Università Roma Sapienza), Annamaria Andreoli (Università di Potenza), Rino Caputo (Università Roma Tor Vergata), Stefano Carrai (Scuola normale - Pisa), Simona Costa (Università Roma Tre), Marco Manotta (Università di Sassari), Clelia Martignoni (Università di Pavia), Michael Rössner (München Universität), Antonio Sichera (Università di Catania).

2. UN MODELLO PER L'EDIZIONE DIGITALE: L'HYPEREDIZIONE DE "IL FU MATTIA PASCAL"

La denominazione di Hyperedizione esprime tutte le potenzialità della nuova *facies* del portale, racchiudendo non solo l'idea del potenziamento e della organizzazione armonica, ma anche quella del 'servizio', nella direzione del lettore e del testo. Scopo dell'edizione digitale è ora quello di favorire la rappresentazione del testo in maniera dinamica (anche nel senso di [4]), senza l'ambizione di esaurirne l'interpretazione, i significati, ma al contrario spingendo il fruitore «verso un ritorno assoluto al testo, al suo godimento, alla sua forza, dispiegabile pienamente solo nel dialogo» [3]. La dimensione plurale del testo come rappresentazione codificata di documenti e informazioni differenti [9] trova quindi una possibile applicazione nell'idea di una 'moltiplicabilità' che si concretizza nelle diverse modalità di lettura, la compenetrazione tra le quali viene favorita da una interfaccia progettata appositamente per l'edizione[2]. Ideata come *hub* principale delle diverse esperienze (ciascuna a propria volta dotata di un'interfaccia dedicata), l'interfaccia dell'Hyperedizione (vd. Fig. 1) non solo permette di realizzare la ricercata scalabilità di consultazione in base al target del fruitore, ma consente anche di incrementare i contenuti riconducibili all'opera in una maniera esponenzialmente maggiore rispetto alle edizioni cartacee e alle DSE tradizionali.

Figura 1. Interfaccia dell'Hyperedizione de Il fu Mattia Pascal

Nello specifico, sono quattro i livelli che vengono messi in connessione tra loro: 1) documentale 2) filologico 3) lessicografico 4) interpretativo.

1) Documentale. La rappresentazione testo-immagine prova a non lasciarsi sedurre da intenti béderiani, mirando invece a rendere le particolarità grafico-scrittorie e correttorie del documento attraverso la rappresentazione delle correzioni. Per *Il fu Mattia Pascal* è stata realizzata un'edizione diplomatica del manoscritto[3] (vd. Fig. 2), il quale si configura come una stesura pressoché in pulito, prossima alla prima pubblicazione del romanzo (in rivista, «Nuova antologia» maggio-luglio 1904) e che testimonia varianti genetiche di notevole interesse linguistico ed ermeneutico [7]. La codifica si serve in particolare dei tag `<del>` e `<add>` per rappresentare le cassature e le aggiunte, con attributi che indicano la tipologia di biffatura, la posizione del testo aggiunto (in interlinea superiore, inferiore, a margine), i tag `<pb>` (con l'attributo `@facs`) per specificare il numero della carta o `<gap>` e `<unclear>` per indicare una lacuna nel testo o nella trascrizione. Nel teiHeader sono invece presentate tutte le informazioni sul manoscritto (msDesc), sulle revisioni (revisionDesc), sul processo editoriale (editorialDecl). L'interfaccia, realizzata con una versione *custom* di EVT1, si apre in una finestra diversa rispetto all'*hub* principale, mentre all'interno dell'interfaccia generale dell'Hyperedizione si possono vedere le carte del manoscritto corrispondenti a ogni pagina del romanzo.

[2] L'interfaccia dell'Hyperedizione è stata progettata da Christian D'Agata.
[3] L'edizione facsimile digitale è stata curata da Christian D'Agata e Alessandro Zammataro. https://www.pirandellonazionale.it/2023/download/evtms/.

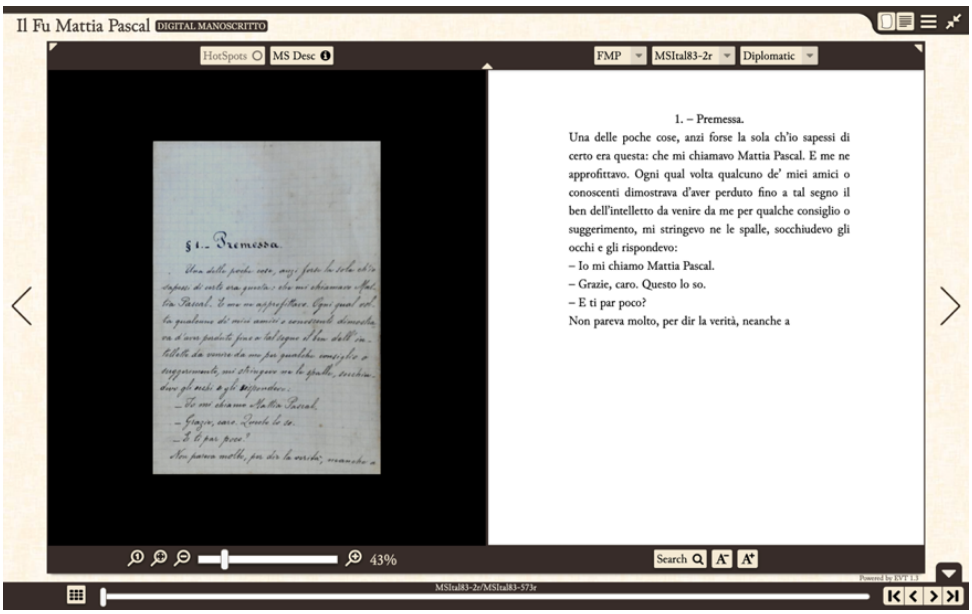


Figura 2. Edizione digitale del ms de Il fu Mattia Pascal *con EVT1*

2) Filologico. È certamente in questo livello che la possibilità di consultazione in maniera scalabile diventa pressoché fondamentale per l'utente. L'affiliazione all'Edizione Nazionale rende necessario preservare l'approccio di tipo tradizionale alla variantistica, che ricalca quello cartaceo attraverso la digitalizzazione PDF di un'edizione statica con apparato critico a piè di pagina[4], certamente il più riconoscibile per il filologo. Così come per l'edizione del manoscritto, anche in questo caso l'*hub* dell'Hyperedizione permette di leggere la pagina dell'edizione statica relativa ai passi del romanzo che si trovano sulla sinistra. Non si poteva però prescindere da un tentativo di restituire vitalità all'inerzia variantistica, attuato attraverso la rappresentazione sinottica dei molteplici testimoni. Questa è stata realizzata con EVT2 e riprende le indicazioni del capitolo 12 (*Critical Apparatus*) delle Guidelines TEI[5] in relazione al Parallel Segmentation Method (e quindi i tag `<app>`, `<lem>`, `<rdg>`)[6]. Si produce così un'edizione capace di sintetizzare una fruibilità avanzata nella visualizzazione con il rigore scientifico nell'allestimento.

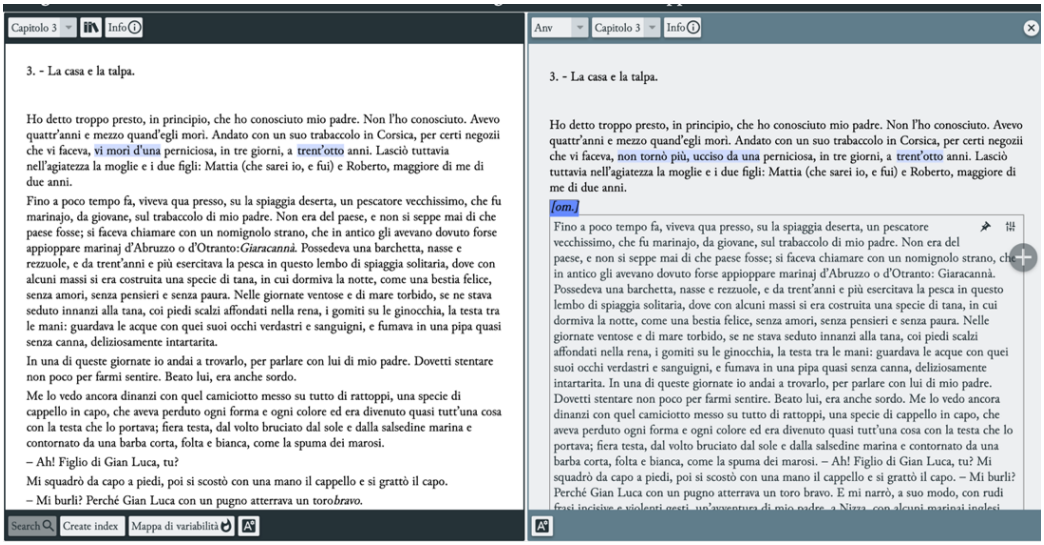


Figura 3. Edizione dinamica de Il fu Mattia Pascal *con EVT2*

3) Lessicografico. Il modo migliore per perfezionare l'interpretazione dei testi letterari è senza dubbio quello di conoscere il vocabolario di uno scrittore, soprattutto se l'indagine viene condotta privilegiando il criterio

[4] L'edizione statica con apparati è a cura di Miryam Grasso. https://www.pirandellonazionale.it/edizione/il-fu-mattia-pascal/
[5] TEI Consortium, eds. "12. Critical Apparatus". *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 4.7.0. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.
[6] L'edizione dinamica è a cura di Giuseppe Canzoneri. https://www.pirandellonazionale.it/2023/download/evt/

intratestuale [10]. Rispetto alla versione originale del portale, l'Hyperedizione de *Il fu Mattia Pascal* prova a potenziare ulteriormente l'interazione tra testo e vocabolario. Per realizzare ciò ci si avvale nel backend di LiotroConcord (un applicativo che si serve dell'ambiente 4D come si può leggere in [3, 5]), all'interno del quale è presente il testo digitalizzato e lemmatizzato. Nell'interfaccia dell'Hyperedizione è dunque sempre possibile interrogare il testo attraverso un form che permette una ricerca per forme o lemmi oppure, in modalità 'vocabolario', selezionando con il mouse una qualunque parola mentre si sta leggendo il testo. Tutte le forme e i lemmi sono così potenzialmente oggetto di una *query* al database che restituisce il contesto KWIC e statistiche sulla frequenza assoluta e relativa.

4) Interpretativo. Grazie alla nuova interfaccia, gli strumenti per l'interpretazione del testo e la sua didattizzazione dialogano con il testo in simultanea, garantendo un'esperienza immersiva realmente multimediale. Il fruitore – immaginato per lo più come uno studente liceale – può interrogare fonti storiche, linguistiche e lessicografiche, visualizzare filmati, ascoltare audioletture senza abbandonare la lettura tradizionale. Infatti, mentre sulla finestra di sinistra si può scorrere il testo del romanzo, su quella di destra si possono consultare le diverse risorse, attivabili anche cliccando sui luoghi del testo direttamente interessati dal commento[7].

In questa nuova veste, *Pirandello Nazionale* si colloca a metà tra edizioni specializzate ed edizioni seriali [6], mirando a una sintesi tra le due proposte. Il sostegno di due importanti istituzioni come l'Università di Catania e la Commissione Nazionale per l'Opera Omnia, nonché quello dell'editore Mondadori (che da subito ha investito nel progetto dell'edizione digitale) ha permesso di farne un laboratorio sperimentale, con un ottimo potenziale attrattivo per studiosi e ricercatori. Al tempo stesso, la scelta di adottare software come EVT, che permettono una visualizzazione immediata della codifica mediante un'interfaccia, semplifica notevolmente il lavoro del filologo, il quale non deve acquisire una particolare competenza nello sviluppo di interfacce web e può limitarsi ad apprendere il linguaggio XML/TEI.

3. L'HYPEREDIZIONE COME STRUMENTO DIDATTICO: UN CASO DI STUDIO

In termini di ricaduta didattica, se i benefici determinati dalla fruizione degli strumenti interpretativi attraverso la nuova interfaccia multimediale sono facilmente prevedibili, più rischioso, in termini di successo, si presentava l'accostamento degli studenti allo sviluppo del portale. Poiché la dimensione seriale che caratterizza (seppur parzialmente) la struttura dell'Hyperedizione rende più rapide e intuitive le operazioni di codifica, si è deciso di coinvolgere alcuni studenti nel lavoro di commento alle opere di Pirandello. Gli studenti chiamati a collaborare frequentano sia Lettere (L-10) e Filologia moderna (LM-14) – quindi due corsi di studio a vocazione tradizionale – per ciò che riguarda più precisamente i problemi testuali, collegati al rapporto tra digitale e letteratura, che quelli di Scienze del Testo per le Professioni digitali (LM-43), coinvolti in particolare per questioni più tecniche come la modellizzazione della codifica. L'occasione per confrontarsi con romanzi e drammi di Pirandello è stata offerta in tutti e tre i casi proposti dalla stesura dell'elaborato finale ed è proseguita poi in forma di tirocinio curriculare presso il CINUM[8]. Le studentesse e gli studenti hanno prima seguito un laboratorio di 6 ore di alfabetizzazione all'XML/TEI, e poi, supervisionati dai docenti relatori e dai tutor, hanno potuto lavorare sul commento ai testi, contribuendo inoltre con le loro esperienze alla riflessione sulla codifica e sulla resa grafica del commento. Prima ancora della concretizzazione di un risultato, l'esperienza si è dimostrata fruttuosa in quanto ha consentito agli studenti di entrare a far parte di una comunità di apprendimento nutrita dall'osmosi continua tra ricerca e didattica e nella quale le distanze generazionali e istituzionali erano pressoché abbattute. Attraverso l'esposizione di tre casi significativi è inoltre possibile comprendere come la riflessione sulla codifica non sia stata condotta in maniera avulsa dal testo, ma al contrario abbia contribuito a migliorarne la comprensione e, di conseguenza, ad affinare le tecniche per il *close reading*.

3.1. Un'Hyperedizione *in fieri*: il caso di *Si gira*

L'intento iniziale della codifica svolta sul romanzo *Si gira* è stato quello di individuare attraverso la codifica XML/TEI l'ossatura base del testo: con il *markup* di `<div>`, `<p>`, `<placeName>`, `<persName>`, `<orgName>`, `<rs>` e `<q>`, si intendeva realizzare una codifica di base che potesse rappresentare gli elementi strutturali del testo, gli attori principali, i luoghi e le interazioni attraverso la codifica del discorso diretto. Tale prima fase ha fatto però emergere sin da subito alcuni degli aspetti più rilevanti su cui condurre l'analisi. Di fronte a elementi testuali complessi come «l'uomo del violino», o

[7] Il commento digitale a *Il fu Mattia Pascal* è un lavoro d'equipe realizzato da Giuseppe Palazzolo, Carmelo Tramontana, Denise Bruno, Elisa Conti e Adriana Damico.

[8] Il riferimento è agli elaborati finali che avevano come oggetto di studio la marcatura di alcuni capitoli del romanzo *I vecchi e giovani*, realizzati da Carola Cunsolo e Gaia Infantino e del romanzo *Si gira!/Quaderni di Serafino Gubbio*, realizzata da Fabrizio Lo Presti.

«la tigre», si è messa in dubbio la concezione stessa di personaggio, come identità definibile da nome e cognome e, di conseguenza, l'uso del tag <persName> è stato esteso, nei casi opportuni, anche a personaggi centrali per l'intera vicenda, seppur anonimi. Discorso simile si può fare per i luoghi, poiché la «Kosmograph» o l'«ospizio di mendicità» non sono solo dei luoghi ma anche organizzazioni di persone: si è dubitato infatti se utilizzare il tag <placeName> o <orgName>. Per quanto riguarda, invece, il tag <rs>, esso è stato utilizzato per identificare tutte le occorrenze di riferimenti indiretti ai personaggi, in questo modo si può comprendere la loro distribuzione nel romanzo e disambiguare sezioni testuali complesse. La marcatura dei dialoghi attraverso <q> ha permesso di strutturare le interazioni verbali e di attribuire ai personaggi un linguaggio (idioletto) e un vocabolario personale (vd. Fig. 4).

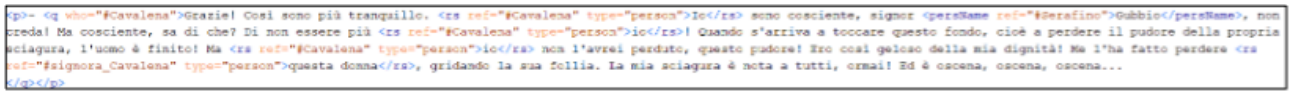


Figura 4. Esempio di marcatura di un dialogo di Si gira

Dopo una prima versione della codifica, nei suoi cicli di revisione, si sono approfondite questioni testuali rilevanti, evidenziabili attraverso una marcatura più avanzata. *Si gira* è un romanzo con un andamento in parte atipico, nel panorama dei romanzi pirandelliani. La struttura apparentemente diaristica, la sua suddivisione in fascicoli e la presenza di sezioni narrative sotto forma di monologo rivolto al lettore, spinge ad una riflessione sulla possibilità di strutturare un secondo livello di codifica che tenga conto proprio di questi aspetti, ad esempio marcando le sezioni monologiche attraverso il tag <seg>. Rimane inoltre aperta la discussione sulla codifica delle occorrenze di termini appartenenti al campo semantico della macchina all'interno del romanzo, che certamente ha come tema centrale proprio il rapporto tra l'uomo, l'arte e la macchina. Inoltre, tra gli elementi emersi dall'analisi testuale condotta su *Si gira* è centrale l'intertestualità nel panorama pirandelliano; tra i vari testi di Pirandello sono ricorrenti infatti apparizioni di parole, espressioni linguistiche, luoghi e in qualche caso personaggi (si vedano gli esempi di "ospizio di mendicità", "puzzo ardente i lavatojo", "ragno nero"). Questi sono solo alcuni dei punti emersi e discussi che meritano di essere indagati nello sviluppo dell'Hyperedizione. È chiaro che l'approccio sia puntualmente funzionale alla ricerca ermeneutica, ed è anche per questo motivo importante sottolineare come gran parte del lavoro sulla codifica sia dettato proprio dalla ricerca delle chiavi di lettura del testo, sempre aperta ad una visione ampia.

3.2. Per un workflow della codifica collaborativa de *I vecchi e i giovani* e dell'*Enrico IV*

Per lo sviluppo dell'Hyperedizione attraverso la collaborazione di laureandi e tirocinanti si è elaborato un workflow che mettesse al centro la rappresentazione in particolare dell'onomastica, della toponomastica, del lessico in disuso, dei collegamenti multimediali, organizzando il lavoro in piccole équipe coordinate da dottorandi e dottori di ricerca, con la supervisione dei docenti e dei direttori dell'edizione digitale. Il workflow è stato improntato sull'utilizzo di Github per lavorare e revisionare collaborativamente la codifica dei testi. La codifica degli elementi base de *I vecchi e i giovani* e dell'*Enrico IV*[9] è stata svolta a partire dal modello sviluppato per *Il fu Mattia Pascal.* In particolare, oltre a <persName>, <placeName>, <orgName>, si è utilizzato il tag <term> per marcare parole, termini, ed espressioni particolari, con definizioni di servizio per il lettore e lo studente, e il tag <objectName> per i cosiddetti "oggetti culturali", ovvero tutti quegli elementi di rilievo in ambito culturale, come libri, opere d'arte, riviste. Similmente si è proceduto a realizzare una codifica di proverbi, modi dire tipici, frasi in lingua straniera, a cui sono state apportate note puntuali. A partire da questa codifica, gli studenti hanno poi in un foglio di lavoro collaborativo inserito le informazioni sui nomi, luoghi, personaggi, link a contenuti multimediali, in modo tale da creare un .csv che con un semplice script Python producesse in maniera automatica il TeiHeader e il back corrispondente (dove inserire l'appendice sul lessico, poi integrato con tutte le informazioni sui responsabili della codifica e delle molteplici revisioni) (vd. Fig. 5).

[9] Il lavoro di codifica sull'*Enrico IV* è stato realizzato da un'équipe composta da studenti e borsisti del Dipartimento di Scienze umanistiche dell'Università di Catania; se ne elencano di seguito i nominativi: Giuseppe Arena, Giulia Cacciatore, Christian D'Agata e Silvia Scuderi.

```
<list type="term">
    <item xml:id="conculcare">
        <term>conculcata</term>
        <gloss>Calpestare o schiacciare deliberatamente qualcosa con i piedi, solitamente in modo aggressivo o distruttivo</gloss>
        <ref target="https://www.treccani.it/vocabolario/conculcare/">Vocabolario Treccani</ref>
    </item>

    <item xml:id="corbelleria">
        <term>corbellerie</term>
        <gloss>Stupidaggine, atto o parole da sciocco, grosso sproposito</gloss>
        <ref target="https://www.treccani.it/vocabolario/corbelleria/">Vocabolario Treccani</ref>
    </item>

    <item xml:id="crocchio">
        <term>crocchio</term>
        <gloss>Gruppo di persone che si riuniscono in cerchio o in un'aggregazione informale per discutere, parlare o svolgere attività comuni.</gloss>
        <ref target="https://www.treccani.it/vocabolario/crocchio2/">Vocabolario Treccani</ref>
    </item>

    <item xml:id="cuora">
        <term>cuora</term>
        <gloss>Strato molle ed erboso che come un prato galleggiante nuota sulle acque di laghi o di paludi</gloss>
        <ref target="https://www.treccani.it/vocabolario/cuora/#:~:text=c%C5%8Frium%20%C2%ABcuoio%2C%20crosta%C2%BB%5D,Navigando%20(D'Annunzio).">Vocabolario Trecc
    </item>

    <item xml:id="fantaccino">
        <term>fantaccini</term>
        <gloss>Soldato semplice di fanteria</gloss>
        <ref target="https://www.treccani.it/vocabolario/fantaccino/">Vocabolario Treccani</ref>
    </item>
```

*Figura 5. Esempio di codifica dell'*Enrico IV

4. CRITICITÀ E IMPLEMENTAZIONI FUTURE

Con l'obiettivo di offrire al lettore strumenti sempre più affinati per navigare all'interno del sito, le prossime tappe saranno orientate a incrementare il commento ai testi e a creare un legame sempre più stretto tra i quattro livelli sopra descritti. Per romanzi come il *Mattia Pascal*, di cui possediamo i testimoni manoscritti, l'obiettivo è creare un collegamento tra gli avantesti e la *princeps*, ovvero tra le varianti scartate e quelle accolte, nonché mettere in risalto la loro valenza semantica nel contesto di approdo. L'interpretazione e trascrizione dei manoscritti, così come la collazione tra le edizioni a stampa, spesso soggette a correzioni e riscritture minuziosamente rilevate nella fascia di apparato delle edizioni critiche, consente di individuare e isolare alcune specifiche scelte linguistiche di Pirandello: attraverso lo strumento delle concordanze, tali scelte verranno esaminate nei contesti lessicali di altre opere e riportati nel commento dell'Hyperedizione in modo da collegare tutte le risorse offerte dal portale (manoscritti, edizioni critiche, concordanze) e farle convergere puntualmente nei testi. Al commento si affiancherà anche una bibliografia critica mirata per approfondire, attraverso studi specifici, alcuni passaggi cruciali dell'opera pirandelliana con l'obiettivo di valorizzare e dare ulteriore scientificità agli strumenti didattici. Una volta collaudati, questi ulteriori strumenti saranno via via estesi alle nuove opere e a quelle già presenti sul sito nell'ottica di sviluppare, quindi valorizzare, il potenziale praticamente illimitato dell'universo pirandelliano.

Come si evince da [3] e dall'aggiornamento testimoniato dal presente contributo, la realizzazione di Pirandello Nazionale ha richiesto e richiede diverse expertise, nonché una quantità di tempo-uomo insostenibile per un singolo ricercatore. La forza del progetto risiede quindi soprattutto nel lavoro di un'équipe composta da studiosi dalle competenze eterogenee e impegnati a vario titolo nello sviluppo del portale. Altresì rilevante il contributo degli studenti, rivelatosi particolarmente fruttuoso, non solo per le ricadute didattiche, ma anche per l'introduzione di un punto di vista diverso (quello del lettore più giovane, colto ma non specialistico), stimolo per nuove riflessioni sulle possibilità del commento digitale.

5. RINGRAZIAMENTI

La pubblicazione è stata realizzata con il cofinanziamento dell'Unione europea - FSE-REACT-EU, PON Ricerca e Innovazione 2014-2020 DM1062/2021.

BIBLIOGRAFIA

[1] Ciotti, Fabio, (a cura di). *Digital Humanities. Metodi, strumenti, saperi*. Roma: Carocci, 2023.

[2] Crucitti, Marilena, Michela Benedetti, Roberta Mirandola, Greta Maneschi, Antonella Soldani, Ludovica Amato, Filippo Lepori, Andrea Taddei, e Federico Boschetti. «La collaborazione inclusiva: un'esperienza didattica di annotazione tramite Euporia». *Umanistica digitale*, fasc. 11 (2021): 145–62. https://doi.org/10.6092/issn.2532-8816/13680.

[3] D'Agata, Christian, Antonio Di Silvestro, e Antonio Sichera. «Edizione critica, edizione digitale, hyperedizione. "Il fu Mattia Pascal" come paradigma dell'Edizione digitale dell'Opera Omnia di Luigi Pirandello». *Bollettino Centro di Studi Filologici e Linguistici siciliani*, fasc. 33 (2022): 263–80.

[4] Del Gratta, Riccardo, Angelo Mario Del Grosso, Simone Zenzaro, Federico Boschetti, e Luigi Bambaci. «La filologia come sistema dinamico». *Umanistica digitale*, fasc. 13 (2022): 1–20. https://doi.org/10.6092/issn.2532-8816/13684.

[5] Giuffrida, Milena, Christian D'Agata, Laura Giurdanella, e Pietro Sichera. «Pirandello Nazionale: per un nuovo modello di edizione digitale, collaborativa e integrata». In *AIUCD 2021 - Book of the extended abstracts*, 207–11. Quaderni di Umanistica Digitale. Umanistica Digitale, 2021. https://doi.org/10.6092/unibo/amsacta/6712.

[6] Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham, Surrey: Ashgate, 2015.

[7] Pirandello, Luigi. *Il fu Mattia Pascal*. (a cura di) Antonio Sichera e Antonio Di Silvestro. Milano: Mondadori, 2023.

[8] Risari, Elisabetta. «Haute couture e prêt-à-porter. Cassola e altri scrittori italiani del secondo Novecento tra Meridiani e Oscar». In *Editori e filologi. Per una filologia editoriale*, (a cura di) Paola Italia e Giorgio Pinotti, 103–10. Roma: Bulzoni, 2014.

[9] Sahle, Patrick. «What is a Scholarly Digital Edition?» In *Digital Scholarly Editing. Theories and Practices*, (a cura di) Matther James Driscoll e Elena Pierazzo, 19–39. Cambridge: Open Book Publishers, 2015.

[10] Tambasco, Itala. «Intratestualità e digitale: prospettive esegetiche sulla 'nuova' filologia dantesca». *Umanistica digitale*, fasc. 14 (2022): 1–17. https://doi.org/10.6092/issn.2532-8816/14959.

[11] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Voci dall'Inferno: Dante per dire il Lager - digitalizzare e studiare le testimonianze

Angelo Mario Del Grosso[1], Marina Riccucci[2], Elvira Mercatanti[3]
[1] CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - angelo.delgrosso@ilc.cnr.it
[2] Università di Pisa, Italia - marina.riccucci@unipi.it
[3] Università di Pisa, Italia - elvira.mercatanti@studenti.unipi.it

ABSTRACT
Il contributo illustra il progetto di ricerca Voci dall'Inferno. L'iniziativa scientifica ha due obiettivi, integrati e correlati: a) la digitalizzazione e la codifica di un corpus, il più ampio possibile di testimonianze non letterarie di sopravvissuti all'olocausto; b) l'individuazione, la quantificazione e la valutazione della presenza di lessico e immagini dantesche all'interno di quelle testimonianze.

PAROLE CHIAVE
Lager; Archivi digitali; Testimonianze Olocausto; XML/TEI; eXist-db.

1. INTRODUZIONE
Nei Lager nazisti, pare oltre 60.000, hanno perso la vita quasi 20 milioni di persone: giovani, vecchi e bambini, uomini, donne, ebrei, dissidenti, zingari e omosessuali, internati militari[1]. Tutto questo è accaduto in un arco di tempo breve: dall'aprile 1943 fino al gennaio 1945. Solo il 10 per cento di coloro che subirono la deportazione nei Lager è sopravvissuto ed è tornato a casa. Di questo 10 per cento circa il 9 per cento è rappresentato da coloro che entrarono in Lager nel 1944. Pochissimi sono i sopravvissuti tra coloro che entrarono in Lager prima di quell'anno: i Lager che noi conosciamo sono i Lager del '44, degli altri sappiamo pochissimo, perché quasi nessuno degli *Häftlinge* è tornato per raccontarlo. Negli ultimi 5 anni è morta la metà dei superstiti. A raccontare la propria esperienza, a dire il Lager a un pubblico che si suppone universale e quindi all'esterno della sfera del proprio privato non sono stati tutti coloro che al Lager sono sopravvissuti, ma una parte, sicuramente non la maggior parte. Di questo gruppo di persone solo la minoranza ha iniziato a farlo subito. Per altri ci sono voluti decenni di impenetrabile, ostinato e doloroso silenzio. Molti sopravvissuti hanno taciuto e quindi la loro storia non la conosceremo mai. Perché il Lager è ineffabile: «Sento talora l'insufficienza dello strumento. Ineffabilità si chiama, ed è una bellissima parola» [9: 688]. Poi, una volta trovata la forza, interviene un altro problema, che è quello di trovare le parole: "Di Auschwitz non si saprà mai tutto, perché alcuni accadimenti sembrano destinati a rimanere senza parole" [23]. Per dire il Lager, dunque, occorre superare una barriera dopo l'altra: quella dell'ineffabilità – riferire significa ricordare e spesso il ricordo di tanta nefandezza è insostenibile – e quella della povertà del vocabolario, un vocabolario che non ha le parole per dirlo, il Lager, un vocabolario senza termini. Dire che il Lager è stato l'inferno [5: 217-220] ha permesso ai sopravvissuti di stabilire un contatto immediato con i loro ascoltatori, con il loro pubblico, con chi non sapeva e non aveva mai voluto sapere[2]. Nelle loro parole ricorre sempre questa metafora condivisa che trova enunciato nella dichiarazione semplice e lineare che il Lager è l'inferno [5: 219]. Sembrerebbe che non ci fosse altro da dire. Dal 2016 Marina Riccucci, docente di Letteratura italiana presso l'Università di Pisa, dirige e coordina, con il supporto del laboratorio CoPhiLab[3] del CNR-ILC[4] di Pisa e del centro di conoscenza CLARIN-IT DiPText-KC[5], il progetto di ricerca Voci dall'Inferno. L'iniziativa scientifica ha due obiettivi, integrati e correlati: a) la digitalizzazione e la codifica di un corpus, il più ampio possibile di testimonianze di sopravvissuti ai Lager; b) l'individuazione, la quantificazione e la valutazione della presenza di lessico e di immagini dantesche all'interno di quelle testimonianze. In questi anni i risultati sono stati sorprendentemente importanti: data la vasta latitudine del progetto (l'altissimo numero di testimonianze, la maggior parte

[1] Per i dettagli sulle statistiche indicate si veda l'enciclopedia dell'Olocausto: https://www.ushmm.org/it
[2] Il saggio "Dante 'per dire' il Lager" [5] riporta espressioni e allusioni dantesche emerse da uno studio dettagliato condotto su un campione di testimonianze di sopravvissuti alle deportazioni; allo stesso contributo si rimanda anche per la letteratura concentrazionaria di riferimento.
[3] https://cophilab.ilc.cnr.it/
[4] https://www.ilc.cnr.it/
[5] https://diptext-kc.clarin-it.it/

delle quali inedite da trattare), è stato necessario coinvolgere metodi e tecniche informatiche e fare della ricerca anche un lavoro collettivo, che ha visto e che vede la collaborazione e il contributo di tanti laureandi[6].

2. IL REPERTORIO DELLE TESTIMONIANZE IN VOCI DALL'INFERNO

Le tipologie testuali attraverso le quali, nei settant'anni e più che hanno attraversato due secoli, il Lager ci è stato restituito "a parole" sono sostanzialmente due: (1) quella della testimonianza diretta – coeva e non – di chi ha vissuto il campo di sterminio e ne ha riferito in forme che solo di rado sono in tangenza con la letterarietà: il *modus dicendi* di questa fattispecie si colloca nello spazio compreso tra il resoconto orale (l'intervista) e quello scritto (il diario, il racconto autobiografico/memoriale, la lettera); (2) quella della testimonianza indiretta – coeva e non – di chi ha vissuto il campo di sterminio e ha scelto, per riferirne, la forma, più spesso della prosa, meno frequentemente della lirica, in ogni caso della narrativa (quindi della letteratura), volendo cioè che il proprio resoconto si presentasse sotto forma di racconto organizzato, tematicamente e stilisticamente strutturato. Appartengono a questo novero le opere (con ambizioni marcatamente) letterarie che mai sarebbero state scritte se non si fossero verificate le determinate contingenze storiche identificabili nell'Olocausto in particolare e nella deportazione in generale. Ci si riferisce, ovviamente, a quell'ampia produzione ai cui vertici sono *Se questo è un uomo* di Primo Levi [10], La notte di Elie Wiesel [23], Da questa parte per il gas di Tadeusz Borowski [3], La specie umana di Robert Antelme [1], Dio è caporale di David Rousset [17], Essere senza destino di Imre Kertész [8]. Nel contesto del lavoro di ricerca chiamiamo questa tipologia di testimonianze "di secondo livello". Siamo di fronte a due forme distinte e differenziate di rappresentazione/restituzione dell'universo concentrazionario. Quanto si è andato progressivamente verificando e acclarando è non solo che Dante rompe il silenzio e interviene nella mente di chi si accinge a narrare la tragedia del proprio nefando vissuto [2, 12, 20, 21] fornendo la parola "che non c'era" - le parole per dirlo, ma anche che, ed è questo il dato più rilevante, a essere influenzate sono le facoltà espressive di tutti i testimoni, anche di chi, magari, Dante lo ha letto, solo e fugacemente, sui banchi di scuola, o, addirittura, anche di chi Dante lo ha solo orecchiato, ricevuto e acquisito come patrimonio di cultura orale. Per un tacito accordo, in nome di una convenzione che tutti i sopravvissuti hanno sottoscritto, e ciascuno autonomamente, ma formando una comunità di fatto, di una parola almeno il vocabolario del Lager, in prima battuta, si compone. Questa parola è Inferno [5: 217, 12]. Tutti – senza eccezione tutti - gli *Häftlinge* che hanno testimoniato, hanno scelto e adottato il lemma Inferno per far capire a chi non lo conosceva e non lo aveva vissuto che cosa fosse il Lager [20: 58]. Ci sono almeno due assunti di base dai quali non si può prescindere: (1) il nesso metaforico (Lager-Inferno) è nesso nuovo, del contemporaneo, del nostro Novecento. Prima non esisteva, semplicemente perché prima degli anni Trenta del Ventesimo secolo, a non esistere era, prima di tutto, il Lager; (2) quando i testimoni parlano del Lager come dell'Inferno non lo fanno riferendosi a un inferno qualsiasi, o a un inferno e basta. Lo fanno avendo in mente l'inferno dantesco [20, 21]; che non sono stati né i poeti né i letterati né gli artisti ad avere abbinato e congiunto, per primi, con la loro competenza tecnica e con la loro alta fantasia, le due sfere, quella dell'inferno dantesco e quella del Lager; certo anche loro, ma insieme e in contemporanea a tutti gli altri che hanno scelto di parlare del Lager e che se la sono sentita di renderne testimonianza e di restituirlo a parole [2]. I dati raccolti fino a questo momento ci dicono che Dante rompe il silenzio, nel senso che interviene sulla mente del sopravvissuto a sciogliere il nodo dell'ineffabilità, a diluire la paralisi della mente e della memoria di fronte all'affiorare del vissuto nefando [5]. La facoltà espressiva, a un certo punto rimanda a un subcosciente reattivo e trova in Dante un motore emancipatore che la mobilita e che produce le parole, in un sussulto vitale e vivifico di resistenza. Ciò accade in tutti i sopravvissuti [15]. Ascoltando e leggendo le testimonianze dei sopravvissuti ai campi di sterminio ci si rende conto che per riferire dell'inferno concentrazionario, gli scampati allo sterminio, persone di ogni livello di istruzione si sono affidate alle parole del vocabolario dantesco, quello della prima cantica, per lo più, quello dell'Inferno [18]. Beninteso, le testimonianze non sono tramate tutte e non sistematicamente o capillarmente di versi danteschi, ma i versi danteschi a un certo punto scoccano dalle labbra di questi uomini e di queste donne, a siglare, a dare la cifra, a esprimere l'inesprimibile. I testimoni possono contare su un patrimonio lemmatico e su un immaginario collettivo fatti di parole dantesche entrate nell'uso, nella vita di tutti i giorni, penetrati nella lingua del quotidiano, dentro il parlare della famiglia e della società, trasmessi di generazione in generazione come un'eredità. L'informatica permette di conservare e di censire, di interrogare e di trovare nessi, di costruire mappe, di intrecciare storie [22].

3. DEFINIZIONE E COSTRUZIONE DELL'ARCHIVIO

Il progetto Voci dall'Inferno ha attraversato nella sua evoluzione tre fasi principali: (1) sviluppo di una banca dati per la gestione dell'anagrafica delle testimonianze - l'archivio ha preso il nome di memoria-archivio [14]; (2) creazione del corpus

[6] Gli studenti che hanno collaborato e collaborano al progetto Voci dall'Inferno sono iscritti al Corso di Studi in Informatica Umanistica e in Italianistica.

delle testimonianze in formato XML/TEI [4]; (3) sviluppo di un applicazione web per la fruizione e l'interrogazione dei dati conservati nell'archivio digitale[7]. La banca dati memoria-archivio, sviluppata da Frida Valecchi, ha consentito di creare un primo inventario delle testimonianze. Quest'ultimo conserva descrizioni catalografiche e letterarie nonché, qualora presenti, gestisce anche le trascrizioni del contenuto testuale della testimonianza. In aggiunta, l'ambiente web permette di confrontare il lessico della testimonianza con il testo della Commedia dantesca da un lato mediante indicizzazione di forme testuali e lemmatizzazione delle stesse, dall'altro mediante confronti di stringhe implementati con politiche di *fuzzy matching* e similarità [11]. L'applicazione consente l'aggiornamento dell'inventario, delle anagrafiche dei testimoni e dei curatori delle fonti. Successivamente, il software ha integrato anche la gestione di documenti in formato XML/TEI. Ad oggi l'archivio dispone di 47 testimonianze di cui 29 trascrizioni solo in full-text e 18 documenti codificati anche in formato XML/TEI.

All'interno dell'archivio digitale le testimonianze si dividono in due macro-classi, le quali ne determinano gli aspetti rappresentazionali, funzionali e fruizionali. Da un lato le testimonianze orali e dall'altro le testimonianze scritte. Pur mantenendo le specifiche differenze, entrambe le classi seguono le indicazioni fornite dalle linee guida del consorzio TEI[8]. In particolare, nel corso del progetto, è stato creato - e via via sempre più raffinato - un "One Document Does it All" (ODD)[9] che dichiara i moduli, gli elementi, gli attributi e i possibili valori ammessi per la codifica del repertorio delle testimonianze. Per quel che concerne il modello di codifica, le testimonianze scritte seguono uno schema di edizione *image-based* [16] di tipo diplomatico-interpretativo con approccio *parallel-transcription*[10] alla rappresentazione del testo-documento [13]. A tale scopo sono stati utilizzati gli elementi definiti nel modulo 11 (Representation of Primary Sources)[11], per la trascrizione della fonte primaria; modulo 13 (Names, Dates, People, and Places)[12], per la rappresentazione delle entità nominate; modulo 16 (Linking, Segmentation, and Alignment)[13], per le analisi di particolari strutture del testo; modulo 17 (Simple Analytic Mechanisms)[14], per l'annotazione semantica e linguistico-lessicale delle unità testuali. Per quanto riguarda la descrizione della fonte primaria, sono stati impiegati, come norma, gli elementi definiti nel modulo 10 (Manuscript Description)[15] delle linee guida TEI.

Le testimonianze orali, di converso, sono caratterizzate da elementi strutturali diversi rispetto al modello di codifica adottato per le testimonianze scritte. In particolare, l'attenzione è rivolta alla dimensione temporale del discorso e all'ordine in cui i vari enunciati si alternano. Ne segue l'interessante definizione di specifici elementi "timeline" finalizzati alla sincronizzazione temporale degli argomenti trattati dal testimone (vd, List. 1a). La sincronizzazione ha lo scopo di allineare il minutaggio degli enunciati con le relative trascrizioni dei parlanti (vd. List. 1b). A complemento della timeline è stata introdotta una sezione di sintesi ragionata, chiamata *regesto* (vd. List. 1a), che illustra brevemente il contenuto di ciascuna divisione. Nel modello di codifica sono state definite quattro diverse timeline. Nella prima timeline gli elementi `<when/>` individuano i momenti in cui si introducono i vari argomenti della testimonianza, quelli cioè riassunti all'interno degli elementi `<item>` presenti nel regesto, i quali, a loro volta, possiedono un attributo `@synch` allo scopo di collegare il minutaggio specificato dal relativo tag `<when/>`. Nella seconda timeline, invece, sono stati raccolti tutti i segmenti in cui le voci si sovrappongono. Per questo, al suo interno, gli elementi `<when/>` sono inseriti a coppie: uno di essi individua l'intervallo della sovrapposizione. La terza timeline è stata invece realizzata per raggruppare gli istanti in cui avviene un cambio di parlante. La quarta ed ultima timeline permette di registrare gli istanti in cui agli enunciati si sovrappongono rumori di sottofondo. Gli elementi XML/TEI più significativi utilizzati per la descrizione delle fonti orali possono essere così sintetizzati: le informazioni relative al supporto sono state registrate mediante l'elemento `<recordingStmt>`, contenuto a sua volta nell'elemento `<sourceDesc>`, appartenente al modulo 8 delle linee guida TEI (Transcriptions of Speech)[16]. L'elemento `<recording>` infine rappresenta una singola registrazione e contiene tutte le informazioni necessarie a specificare il contesto e le responsabilità della registrazione. Ogni elemento `<recording>` è accompagnato dall'attributo di tipo (`@type`) per specificare la natura audio o video, e di una durata (`@dur`) per ogni singola registrazione.

[7] Tra le iniziative simili si citano Let Them Speak, https://lts.fortunoff.library.yale.edu/, oppure Boder: from wire recordings, https://ranke2.uni.lu/u/boder/, oppure archivi quali CDEC, https://digital-library.cdec.it/cdec-web/, ed iniziative come EHRI, https://www.ehri-project.eu/. Altri progetti simili sono indicizzati dal progetto https://dhjewish.org/projects.

[8] https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

[9] Lo schema di codifica in formato ODD è attualmente in fase di revisione e sarà disponibile sul repository github del progetto assieme alle altre risorse aperte: https://github.com/CoPhi/voci-inferno.

[10] https://tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PH-bov

[11] https://tei-c.org/release/doc/tei-p5-doc/en/html/PH.html

[12] https://tei-c.org/release/doc/tei-p5-doc/en/html/ND.html

[13] https://tei-c.org/release/doc/tei-p5-doc/en/html/SA.html

[14] https://tei-c.org/release/doc/tei-p5-doc/en/html/AI.html

[15] https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html

[16] https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

Infine, all'interno del blocco <profileDesc> sono stati impiegati gli elementi definiti nel modulo 15 (Language Corpora)[17], in particolare l'elemento <particDesc> offre una descrizione accurata delle persone che hanno preso parte alla conversazione.

```
<abstract><ab><list>
<item synch="#TR1">
<persName ref="#LS">
<forename>Liliana</forename>
<surname>Segre</surname>
</persName>, rispondendo
alla domanda postale da <persName ref="#AS">
<forename>Anna</forename>
<surname>Segre</surname>
</persName>, parla di che cosa abbia voluto
significare andare …</item></abstract>
```

Listato 1a. Esempio regesto (Liliana Segre)

```
<standOff>
<timeline xml:id="TL1I"
source="#reg_1B" unit="s">
<!-- … -->
<when xml:id="TR1"
  absolute="00:00:00"/>
<!-- … -->
<when xml:id="TR7"
absolute="00:23:41"/>
<!-- … -->
</timeline><!-- … --></standOff>
```

Listato 1b. Esempio timeline (Liliana Segre)

I due modelli di codifica introdotti, vale a dire il modello per le testimonianze orali e quello per le testimonianze scritte, si differenziano tra loro sia per le scelte descrittive sia per quelle analitiche. La struttura logica della testimonianza scritta segue spesso una grammatica prevalentemente epistolare, ma può differire sostanzialmente per caratteristiche autoriali e redazionali (interventi autoriali su manoscritti oppure dattiloscritti). La rappresentazione della fonte primaria fa uso del tagset *facsimile* con il quale descrivere il perimetro delle aree di interesse mediante opportuni attributi presenti nell'elemento @zone. Le aree saranno poi riferite dagli elementi corrispondenti collocati nella sezione riservata alla codifica della trascrizione, nel rispetto delle buone prassi suggerite dal metodo parallel-transcription (modulo 11 TEI). Nella sezione dedicata alla trascrizione (blocco <text>) saranno quindi presenti elementi destinati alla rappresentazione del contenuto della fonte primaria, quali ad esempio errori ortografici oppure errori di battitura, codificati con l'elemento <subst> oppure con l'elemento <mod>. Dove presenti danni materiali l'elemento <damage> e l'elemento <choice> con le relative articolazioni sono usati per registrare la lezione dell'originale e la lettura mediata da una più onerosa interpretazione del trascrittore/editore, che trova, qualora necessario, nell'elemento <supplied> la possibilità di integrazioni e supplementi. Le entità nominate seguono le prassi suggerite dalle linee guida adottando gli elementi <person>, <org>, <place>, <event> e i rispettivi <personName>, <orgName>, <placeName>. Le citazioni e la terminologia dantesca sono state annotate con gli elementi <cit> e <term> rispettivamente . Per quanto riguarda invece le testimonianze orali la trascrizione viene divisa in unità testuali dette utterances (enunciati) mediante l'uso dell'elemento <u> (vd. List. 2a). Ogni enunciato è accompagnato dall'attributo @who che permette di associare ad esso la persona che lo ha formulato. In aggiunta, gli elementi @xml:id e @synch sono funzionali ad una corretta sincronizzazione con le timeline discusse in precedenza. L'attributo @trans, invece, specifica se gli enunciati dei partecipanti si susseguono oppure si sovrappongono. Nel corso della testimonianza, numerosi fenomeni sono registrati, quali pause (elemento <pause> con attributo @type per indicare la lunghezza), suoni non lessicali (elemento <vocal>), eventi prossemici (elemento <kinesic>), rumori di sottofondo (elemento <incident>), passi inudibili oppure incerti (elementi <gap> e <unclear>), cambiamenti di caratteristiche paralinguistiche quali intonazione, volume, ritmo, velocità mediante elemento <shift/> accompagnato come si conviene dagli attributi @feature e @new (vd. List.2a e 2b).

```
<u><!-- … --!>
non hai un nome, perché
hai un numero, <pause type="long"/> ti
chiamano per numero
<pause type="medium"/> e quindi <pause
type="short"/> cercano di
degradarti <pause type="short"/> con la
fame <pause type="short"/>
<!-- … ---!></u>
```

Listato 2a. Esempio pausa

```
<u who="#MARCHERIA" xml:id="m223"
synch="#tlp457"> In questo
<supplied>caos</supplied>
sì, perché arrivavano i russi <pause
type="short"/>
e c'era il caos. Siamo <del
type="repetition">siamo</del>
 <kinesic> <desc>Ida mostra la grandezza
della piazza</desc></kinesic>
```

Listato 2b. Esempio altri fenomeni

[17] https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html

4. L'APPLICAZIONE VOCI DALL'INFERNO

L'archivio digitale delle testimonianze sarebbe meno efficace dal punto di vista funzionale e scientifico senza la presenza di un componente software dedicato all'estrazione, alla manipolazione, alla presentazione e alla fruizione dei dati prodotti durante la fase di codifica. Nel corso del progetto sono state sperimentate due differenti strategie di restituzione dei dati che si avvalgono rispettivamente di due differenti approcci architetturali. Per il primo approccio sono state sviluppate applicazioni di fruizione web facendo leva sulle funzionalità di una libreria *javascript* client-side per l'elaborazione di documenti in formato XML: *SaxonJS2*[18]. Relativamente al secondo approccio, invece, le applicazioni web sono state sviluppate in ambiente eXist-db[19] mediante l'uso del modulo *HTML templating* (server-side). Grazie all'uso della libreria SaxonJS2 è possibile integrare un efficiente processore XSLT demandando al browser la manipolazione dell'oggetto DOM della pagina HTML con i dati recuperati dal documento XML. La libreria espone una ricca API i cui metodi principali sono *SaxonJS.transform (options[,execution])* per eseguire le istruzioni definite dalle regole di trasformazione del foglio di stile e *SaxonJS.XPath.evaluate(XPath, contextItem?, options?)* per selezionare opportune sequenze di nodi XML oppure elaborare frammenti XML secondo le specifiche dello standard XPath 3.1 (vd. List. 3).

```
SaxonJS.getResource({ location: "testimone.xml", type: "xml" }).then(doc => { const
result = SaxonJS.XPath.evaluate("//persName/text()", doc); const output =
SaxonJS.serialize(result, {method: "xml", indent: true,"omit-xml-declaration":true}); })
```

Listato 3. Estratto codice SaxonJS2 per uso con espressione XPath

L'immagine in Figura 1 mostra una pagina Web generata con l'ausilio della libreria SaxonJS2 per la visualizzazione della testimonianza di Arminio Wachsberger. È possibile notare il testo trascritto dell'intervista, i partecipanti, i fenomeni testuali annotati e resi graficamente secondo gli stili indicati in legenda.

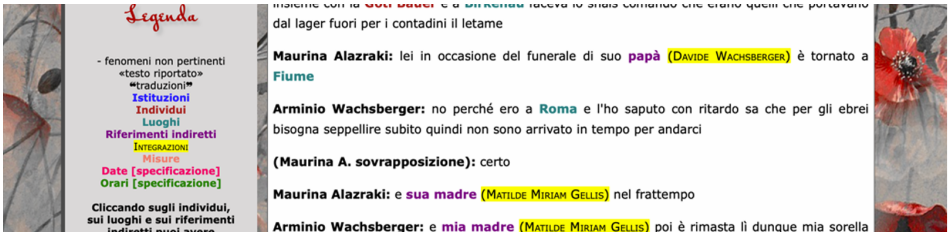


Figura 1. Pagina HTML di un estratto della testimonianza di Arminio Wachsberger

Il secondo approccio è quello basato sulla tecnologia disponibile per il database eXist-db. Come introdotto, la piattaforma integra un modulo dedicato alla generazione dinamica di pagine HTML a partire da collezioni di documenti in formato XML e da funzioni implementate mediante il linguaggio di interrogazione XQuery[20]. Il funzionamento di base prevede l'uso di documenti-modello in HTML (template), in cui si aggiungono opportune direttive e chiamate a funzioni XQuery. Le funzioni implementano la logica applicativa dedicata alla generazione dei frammenti HTML utili a completare l'effettiva pagina visualizzata dal browser. Una caratteristica rilevante della tecnologia eXist-db è la possibilità di avvalersi della libreria Apache Lucene[21] per l'indicizzazione dei dati testuali e per la conseguente interrogazione degli stessi. La Figura 2 mostra un esempio di interrogazione e restituzione dei dati relativa alla testimonianza di Ida Marcheria (ricerca parola parziale "bell").

[18] https://www.saxonica.com/download/javascript.xml
[19] https://exist-db.org/exist/apps/homepage/index.html
[20] https://www.w3.org/TR/xquery-31/
[21] https://lucene.apache.org/

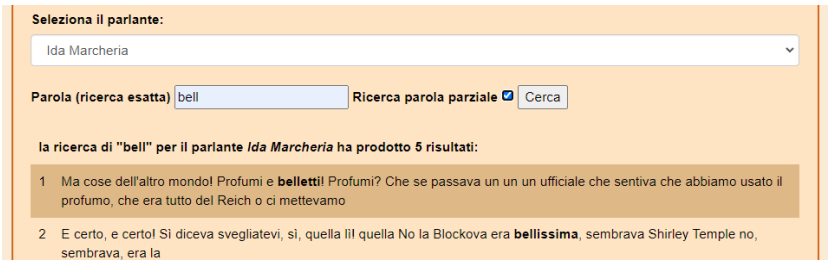


Figura 2. Ricerca per parola parziale implementata per la testimonianza di Ida Marcheria

L'applicazione web realizzata fino ad oggi per Voci dall'Inferno ha molteplici funzionalità implementate o in corso di sviluppo (vd. Figg. 3a e 3b) quali: (a) gestione e ricerca in catalogo; (b) presentazione e fruizione dei dati in parallelo con la fonte primaria; (c) ricerca all'interno dell'archivio testuale; (d) gestione del regesto; (d) gestione delle convenzioni del parlato [7]; (e) statistiche dei fenomeni; (f) gestione della terminologia; (g) gestione delle citazioni e delle allusioni (dantesche in particolare). Tra le funzionalità in sviluppo quelle più importanti sono dedicata alla classificazione ed estrazione dei dati secondo tecniche di machine-learning per la trascrizione automatica del parlato, la ricerca automatica di tasselli letterari nonché tecniche automatiche di network-analysis [19].

Figura 3a. Ricerca in Catalogo

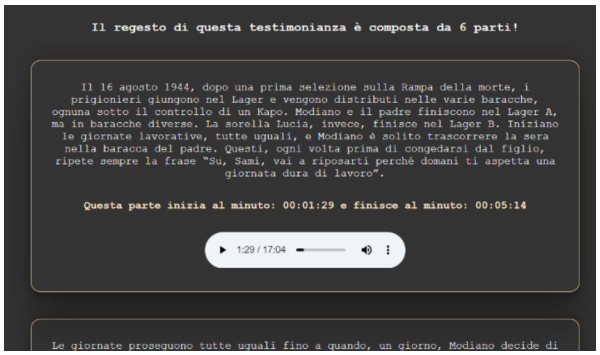


Figura 3b. Regesto e ascolto originale del testimone

BIBLIOGRAFIA

[1] Antelme, Robert. *The Human Race*. Tradotto da Jeffrey Haight e Annie Mahler. Northwestern: Marlboro Press, 1998.

[2] Arquès, Rossend. «Dante nell'inferno moderno: la letteratura dopo Auschwitz». *Rassegna Europea di Letteratura Italiana* 33 (2009): 89–110.

[3] Borowski, Tadeusz. *This way for the gas, ladies and gentlemen*. New York: Penguin Books, 1976.

[4] Burnard, Lou. *What Is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources*. Marseille: OpenEdition Press, 2014.

[5] Calderini, Sara, e Marina Riccucci. «L'ineffabilità della nefandezza: Dante "per dire" il Lager: un sondaggio preliminare nelle testimonianze non letterarie». *Italianistica* 49 (2020): 213–28. https://doi.org/10.19272/202001301011.

[6] Devlin, Jacob, Chang Ming-Wei, Lee Kenton, e Kristina Toutanova. «BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding». *arXiv*, 2019. http://arxiv.org/abs/1810.04805.

[7] Ehlich, Konrad. «HIAT: A Transcription System for Discourse Data». In *Talking Data: Transcription and Coding in Discourse Research*, (a cura di) Jane Edwards e Martin D. Lampert, 123–48. Hillsdale, NJ: Erlbaum, 1993.

[8] Kertész, Imre. *Fatelessness*. Vintage, 1975.

[9] Levi, Primo. *Opere Complete*. (a cura di) Marco Belpoliti. Giulio Einaudi editore, 2018.

[10] Levi, Primo. *Se questo è un uomo*. Torino: De Silva, 1947.

[11] Moritz, Maria, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, e Marco Büchler. «Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and Its Application to Bible Reuse». In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1849–59. Austin, Texas: Association for Computational Linguistics, 2016. https://doi.org/10.18653/v1/D16-1190.

[12] Pertile, Lino. «L'inferno, il Lager, la poesia». *Dante: Rivista internazionale di studi su Dante Alighieri* 7 (2010): 11–34. https://doi.org/10.1400/166873.

[13] Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham, Surrey: Ashgate, 2015.

[14] Riccucci, Marina, Angelo Mario Del Grosso, Frida Valecchi, e Giulia Causarano. «Testimoniare Il Lager: l'informatica al servizio della memoria». In *AIUCD 2021 - Book of the extended abstracts*, 567–72. Quaderni di Umanistica Digitale. Umanistica Digitale, 2021. https://doi.org/10.6092/unibo/amsacta/6712.

[15] Riccucci, Marina, e Laura Ricotti. *Il dovere della parola. Le testimonianze di Liliana Segre e di Goti Herskovits Bauer*. Pisa: Pacini Editore, 2021.

[16] Robinson, Peter. «Towards a Theory of Digital Editions». *Variants*, fasc. 10 (2013): 105–31.

[17] Rousset, David. *L'Univers concentrationnaire*. Paris: Éditions du Pavois, 1946.

[18] Segre, Anna, e Gloria Pavoncello. *Judenrampe. Gli ultimi testimoni*. Roma: Elliot Edizioni, 2010.

[19] Suissa, Omri, Elmalech Avshalom, e Maayan Zhitomirsky-Geffet. «Text Analysis Using Deep Neural Networks in Digital Humanities and Information Science». *Journal of the Association for Information Science and Technology* 73, fasc. 2 (2022): 268–87. https://doi.org/10.1002/asi.24544.

[20] Sustersic, Federica. «La dicibilità del male. La ricezione dantesca nelle testimonianze concentrazionarie». *Dante: Rivista internazionale di studi su Dante Alighieri* 13 (2016): 57–78. https://doi.org/10.19272/201605401005.

[21] Taterka, Thomas. *Dante Deutsch. Studi sulla letteratura dei Lager*. Tradotto da Enrico Paventi. Viterbo: Sette Città, 2002.

[22] Vitali, Giovanni Pietro. «Visualizing Second World War Violence through an Atlas of Nazi–Fascist Repression». *Digital Scholarship in the Humanities* 37, fasc. 2 (2021): 594–610. https://doi.org/10.1093/llc/fqab070.

[23] Wiesel, Elie. *La notte*. Tradotto da Vogelmann, Daniel. Collana di narrativa De Agostini per la scuola. Giuntina editore, 1995.

DIZIONARI E DIGITALIZZAZIONE DI BANCHE DATI

Il VIVer (Vocabolario dell'Italiano Verista)

Gabriella Alfieri[1], Marco Biffi[2], Stephanie Cerruto[3], Giovanni Salucci[4]
[1] Università di Catania, Italia – alfieri@unict.it
[2] Università di Firenze, Italia – marco.biffi@unifi.it
[3] Fondazione Verga, Italia – stephaniecerruto@outlook.com
[4] Università di Firenze, Italia – giovanni.salucci@unifi.it

ABSTRACT

Il progetto del VIVer punta sulla lessicografia informatica per rinnovare il panorama storico-letterario e storico-linguistico, rientrando altresì nell'intento programmatico della Fondazione Verga di rivisitare il verismo nel quadro sovranazionale del realismo letterario. Più in generale il VIVer mira a descrivere e ridefinire i linguaggi del realismo italiano e a incrementare la conoscenza dell'italiano postunitario, con settori semantici finora poco esplorati. A partire da un corpus multigenere (narrativa, teatro, letteratura educativa e testualità metadiegetica) si punta a costruire un vocabolario digitale che, sulla scia del VoDIM (Vocabolario Dinamico dell'Italiano Moderno) consultabile sulla Stazione lessicografica dell'Accademia della Crusca, restituisca nella sua organicità l'italiano verista nelle sue componenti essenziali: lessico regionalizzato, fraseologia, tecnicismi socio-ambientali e metadiegetici. Il corpus è digitalizzato tramite procedure di OCR, revisionato e immesso in una piattaforma sviluppata per il progetto. I testi sono poi marcati secondo gli standard XML/TEI per rilevare il repertorio categoriale che spazia dai regionalismi ai proverbi e alle sentenze. Attualmente la banca dati consiste in oltre 200 testi, di cui 33 testi già revisionati, in corso di marcatura, e caricati su una prima versione prototipale del sito[1]. I principali risultati attesi sono: costituire una Sala di lettura ad accesso libero che contenga un corpus dei corpora della letteratura verista; realizzare un vocabolario dinamico dell'italiano letterario post-unitario a consultazione variabile per fini di ricerca e di didattica; offrire una prima descrizione lessicografica della fraseologia dell'italiano moderno con particolare riferimento alle specificità del verismo (es. codice gestuale).

PAROLE CHIAVE

Vocabulary of Verism; Phraseology; Database; Metadata; XML/TEI.

1. INTRODUZIONE

La Fondazione Verga, in collaborazione con l'Accademia della Crusca, sta realizzando dal 2017 il VIVer "Vocabolario dell'Italiano Verista" basato su un corpus multigenere rappresentativo dell'esperienza verista in tutti i suoi aspetti. Per assicurare un avanzamento costante del progetto, nel 2023 è stata costituita una rete di 14 università e centri di ricerca che sta procedendo, a seconda dell'area di appartenenza degli autori e delle autrici, alla ricognizione dei testi e alla marcatura, e che successivamente si occuperà anche della lemmatizzazione.

Il progetto mira a colmare una vistosa lacuna nella storia dell'italiano letterario e dell'italiano contemporaneo in generale: la conoscenza su base descrittiva dell'italiano letterario post-manzoniano, rappresentato da narrativa, teatro e testualità educativa e metadiegetica del periodo postunitario. L'apporto di Verga e dei veristi al rinnovamento e all'arricchimento dell'italiano post-manzoniano è stato sancito da vari studi teorici, che attendono tuttora un supporto descrittivo [1, 2, 11]. Al drastico monocentrismo fiorentino poi adottato dall'autorità governativa si contrapponeva un'unificazione linguistica 'morbida', fondata sulla progressiva convergenza 'federativa' ma toscano-centrica di esperienze letterarie, filologiche e lessicografiche di origine regionale [15]. In prospettiva il VIVer contribuirà a verificare l'effettivo tasso di convergenza idiomatica nell'italiano contemporaneo, costituendo per gli storici della lingua una fonte ineludibile per ricostruire e descrivere un patrimonio lessicale e fraseologico ancora in gran parte ignorato o sottovalutato. Il progetto si pone come obiettivo di rendere accessibile nel suo complesso la testualità postunitaria, e verista in particolare, almeno a tre tipologie di destinatari: a) comunità scientifica con finalità di avanzamento delle conoscenze in ambito filologico-critico, storico-letterario e storico-linguistico; b) comunità dei lettori con finalità di divulgazione culturale; c) comunità scolastica a fini di azione didattica.

2. CORPUS

Il corpus abbraccia un orizzonte geoletterario e geolinguisitico variegato che, a partire dalla Sicilia, centro di irradiazione della testualità verista, si estende da Piemonte, Lombardia, Veneto e Friuli, attraverso Toscana, Abruzzo, Campania,

[1] Consultabile all'indirizzo https://testi.progettoviver.it/.

Lucania e Calabria, agli ambiti insulari della Sardegna, e anche di Corsica e Malta in quanto aree italofone all'epoca del verismo. Il progetto si fonda su solide basi filologiche, assumendo come testi di riferimento edizioni critiche o autorevoli edizioni commentate o, in assenza di queste, le prime edizioni. Il corpus, sulla base dei più attuali e autorevoli studi sul verismo [3, 9, 10], abbraccia il periodo 1850-1922, in un'accezione estensiva di "Verismo" come declinazione italiana del Realismo e Naturalismo europeo (da Berthold Auerbach, a Thomas Hardy a Èmile Zola), nonché come attività intellettuale inclusiva di molteplici generi testuali. Attualmente il corpus include più di 200 testi ed è articolato nei seguenti generi testuali: 1. narrativa (novelle, romanzi di autori e autrici rappresentativi del realismo e del verismo); 2. poesia verista; 3. teatro; 4. realismo sociale (autori e autrici che scrivevano di questioni sociali del tempo – autori a metà fra realismo e verismo); 5. verismo documentario (giornalismo – inchieste); 6. testi teorici e manifesti poetici (es. tutte le prefazioni degli autori alle raccolte di novelle o romanzi; recensioni di critici letterari). In futuro il corpus verrà ampliato con l'inclusione degli epistolari e delle traduzioni italiane apparse tra Otto e Novecento di Zola, Auerbach, Hardy e di altri autori rappresentativi del realismo europeo.

La fase di trascrizione del corpus è stata realizzata attraverso metodologie e soluzioni di OCR open source (Tesseract[2]) e commerciali (ABBYY Finereader, Nuance Omnipage), scegliendo il formato plain text al fine di preparare i testi a una successiva fase di revisione. I documenti in formato testo sono stati poi inseriti nella piattaforma creata per il progetto e, come si dirà più approfonditamente dopo, codificati secondo una marcatura di tipo strutturale e linguistico.

3. METODOLOGIA E CODIFICA

La costruzione di un vocabolario "reticolare" come il VIVer ha inglobato non solo i processi informatici tipici di una digitalizzazione ma anche uno studio informatico-linguistico capace di soddisfare gli obiettivi che il progetto si propone di perseguire. Il portale del VIVer, come da definizione, prevede contenuti e servizi nativi, accesso a risorse esterne e personalizzazione per gli utenti. Per quanto l'obiettivo specifico del progetto sia la realizzazione di un dizionario, il risultato finale prevede anche la presenza di una banca dati testuale verista, che non solo costituirà il corpus rappresentativo di partenza per i lessicografi che redigeranno le voci, ma offrirà essa stessa uno strumento di consultazione di alto livello che permetterà, oltre alle indagini di tipo lessicale, in certa misura anche quelle di tipo grafico, morfologico e sintattico.

Il primo aspetto a cui si è prestata particolare attenzione è stata l'architettura della banca dati, che è fondamentale per riuscire a ottenere le informazioni necessarie alla ricerca. La banca dati del VIVer è organizzata su quattro livelli: 1) corpus generale; 2) sottocorpus specifico; 3) testo; 4) unità testuale. La divisione in quattro livelli è fondamentale per poter gestire in modo adeguato i risultati delle ricerche. Una specifica attenzione merita il secondo livello, quello dei sottocorpora, che si intende gestire in modo dinamico, assegnando a ciascun testo più classificatori (di genere, di tipologia testuale, ecc.) in modo da consentire all'utente di circoscrivere, in fase di futura interrogazione, gruppi di testo specifici in base alle proprie esigenze. Va sottolineato, invece, che il quarto livello è più funzionale a una buona gestione del testo per consentire alla procedura informatica di interrogazione di raffinare in modo opportuno la ricerca.

L'utilizzo di una piattaforma web (vd. Fig. 1) condivisa per la gestione e marcatura dei testi, ma profilata nei permessi di lavoro, è un elemento di particolare importanza in un progetto che prevede l'intervento di numerosi gruppi di lavoro, ciascuno con diversi operatori che si alternano nelle varie fasi di lavoro. Per la realizzazione informatica del Corpus è stata scelta la piattaforma WCM di Progettinrete, già utilizzata in altri progetti letterari e linguistici (il portale Carte d'autore, il VODIM, l'Atlante gastronomico della lingua italiana).

I testi sono marcati secondo gli standard XML/TEI. Il sistema che si è messo a punto prevede una piattaforma di *back office* funzionale a una elevata operatività dei redattori. A seguito della trascrizione dei testi e della loro revisione, si procede in piattaforma con la creazione delle schede Opera, associando per ciascuna i principali metadati; successivamente sono caricati in piattaforma i vari testi, organizzati per parti (oggetti). La piattaforma consente anche di acquisire testi premarcati con editor XML, sui quali vengono eseguiti automaticamente controlli di correttezza formale e di adeguatezza degli standard di caratteri (eventualmente ricondotti ad omogeneità per favorire le procedure di tokenizzazione, indicizzazione e interrogazione). Nonostante tale possibilità, la procedura naturale e ideale è invece quella dell'inserimento del testo non ancora lavorato, acquisito e collazionato, che, dopo essere stato sottoposto al controllo dei caratteri e alle procedure di omogeneizzazione, è reso disponibile per la marcatura all'interno della piattaforma, nel corretto livello della struttura.

Nella piattaforma è configurata una serie di pulsanti personalizzati in funzione dei tag specifici della DTD del progetto e quindi la marcatura si svolge in una modalità estremamente funzionale. Questo approccio è certamente quello che garantisce la massima omogeneità e organicità del testo, della sua marcatura, della sua architettura, rispetto ad altre soluzioni che prevedano l'utilizzo di programmi locali.

[2] https://github.com/tesseract-ocr/tesseract

Le modalità di marcatura sono estremamente semplici e intuitive. I tag della testata (l'*header*) del documento XML/TEI sono "tradotti" nella forma di un campo di database e liberano il redattore dalla fatica di una marcatura per evidenziatura che è a tutt'oggi la più efficace strategia dei principali editor XML. Nell'*header* sono previste marcature essenziali all'identificazione della fonte e alla sua datazione (titolo, autore, editore, luogo di edizione, anno). È naturalmente l'*header* la zona interessata all'introduzione di classificatori che poi consentano la generazione di corpora personalizzati in funzione di specifiche ricerche (genere, tipologia, livello diastratico del pubblico di riferimento ecc.). Ed è per sua stessa natura la zona più elastica, pronta ad accogliere eventuali marcature aggiuntive in fase di elaborazione del corpus (le modifiche "retroattive" sui testi già marcati sono infatti decisamente limitate nel numero). Nell'*header* sono registrati anche i metadati amministrativi, vale a dire quelli che contengono informazioni su licenze, copyright e quindi specifiche sulla possibilità di riutilizzo dell'edizione digitale (in questo modo ogni file in formato XML marcato sarà quindi di fatto un'edizione digitale dell'opera originaria). In piattaforma il TEIHEADER verrà ricavato dai dati e metadati di schedatura. Gli oggetti invece dovranno coprire la parte di TEXT. La marcatura prevede due livelli, quello strutturale e quello linguistico. Dato l'interesse per la costruzione del corpus in vista della interrogazione ai fini della costruzione del Vocabolario, si è optato per una leggera marcatura strutturale, identificando quindi non solamente le porzioni di testo previste dalla struttura XML/TEI, ma anche marcando parti di testo non rilevanti, oppure molto rilevanti. Per motivi di economicità non è stata svolta nessuna marcatura presentazionale dei testi, fatta eccezione per il cambio verso nei brani di poesia presenti; sono invece stati marcati tutti i cambi pagina, per poter gestire l'eventuale aggancio con le immagini in facsimile (laddove disponibili) e ricostruire la collocazione dei vari lemmi o forme rintracciate. I marcatori linguistici, invece, consentono di far emergere le specificità linguistiche del verismo nel senso sopra prospettato, che, sulla scorta di importanti studi descrittivi [8], nonché in seguito a varie sperimentazioni con livelli di maggiore o minore granularità, sono state infine identificate in tre macro-aree: regionalismi (compresi i toscanismi e i dialettismi, `<distinct type="regionalismo">`), tecnicismi (sia quelli socio-ambientali sia quelli meta-diegetici, `<distinct type="tecnicismo">`) e formularità (tutte le strutture fraseologiche, a partire dalle similitudini convenzionali fino ad arrivare ai proverbi e alle sentenze, `<distinct type="formularita">`). In corso d'opera si è poi deciso di marcare anche le neoformazioni d'autore e le retrodatazioni. Infine, utilizzando un tag `<span>` viene marcata la forma ricostruita di espressioni multiparola i cui componenti possono trovarsi anche distanziati nel testo.

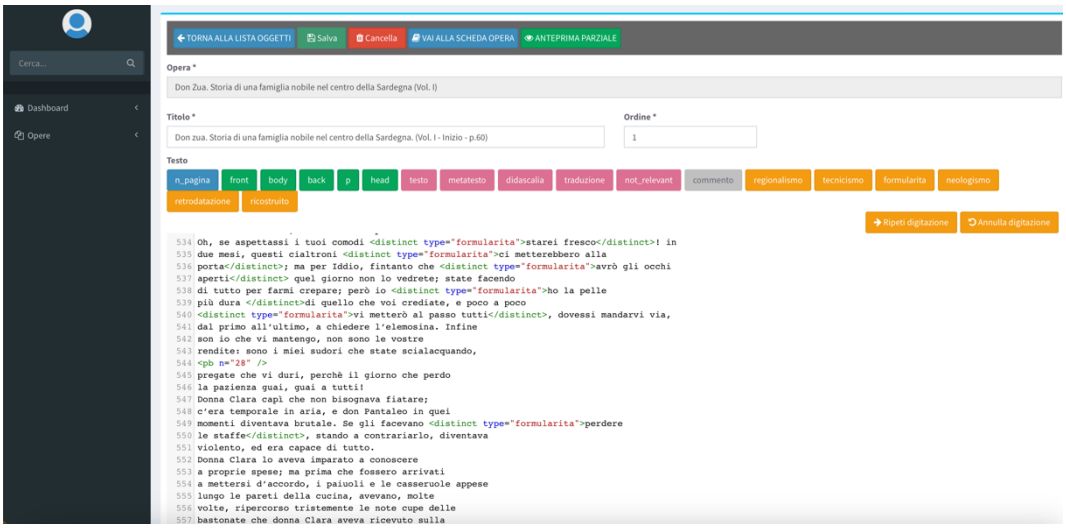


Figura 1. Piattaforma di marcatura del VIVer - Esempio di marcatura linguistica

4. VALORIZZAZIONE E TRASFERIMENTO DELLA CONOSCENZA

Il VIVer si configura come un cantiere lessicografico digitale in cui le attività scientifiche hanno una forte integrazione con quelle informatiche. Nei progetti lessicografici digitali viene consigliato di utilizzare gli Identificatori persistenti [14] e si è quindi deciso di utilizzare il DOI (Digital Object Identifier) come strumento per la diffusione e promozione, scegliendo Crossref come agenzia di registrazione [13].

Oltre ad assegnare e registrare un DOI per ogni edizione digitale delle nuove opere digitali in formato XML, si è scelto di registrare una serie di DOI anche per la banca dati del Vocabolario, in corrispondenza dei vari livelli gerarchici con cui il database del VIVer è organizzato (*database*, *collection*, *record*, quest'ultimo corrispondente alla scheda del dizionario). Si prevede la compilazione di metadati di qualità e ricchi di informazioni, con l'obiettivo di favorire la diffusione del progetto

e di massimizzarne l'impatto, attraverso l'indicizzazione dei dati e metadati negli aggregatori accademici e nei motori di ricerca generalisti.
Tenuto conto delle specificità del progetto, che oltre alla dimensione scientifica e lessicografica punta a un allargamento a un pubblico più ampio, nella scelta dei metadati si è optato per l'adozione di profili sia specialistici che generali. Innanzi tutto, è previsto l'utilizzo del Dublin Core; oltre alla presenza del profilo CrossRef (per la registrazione del DOI secondo il tracciato *dataset*) si è deciso per l'utilizzo di Highwire Press per favorire l'indicizzazione in Google Scholar e negli aggregatori accademici. Infine, per favorire la diffusione e la condivisione nelle reti social, si adotta il profilo Open Graph[3] e Twitter:Cards[4]. Scelte simili sono già state effettuate in progetti analoghi, e hanno avuto un immediato e positivo riscontro nell'esito della indicizzazione su Google e sugli altri motori di ricerca, sia generalisti che accademici: già pochi giorni dopo il rilascio pubblico, le pagine sono state indicizzate e risultavano presenti nelle prime posizioni dei risultati per le parole chiave pertinenti.

5. PRIMI RISULTATI E PROSPETTIVE ATTESE

La piattaforma di interrogazione della banca dati testuale è in fase di controllo e di implementazione del prototipo. Prevede un motore di ricerca per forma che consente l'impiego dei caratteri jolly, degli operatori booleani (anche a distanza fissata dall'utente), la ricerca espansa (senza tener conto degli accenti); e dovrà integrare strumenti di analisi statistica (in particolare in relazione alle cooccorrenze, per facilitare l'individuazione di espressioni idiomatiche anche nascoste). Ma si sta valutando anche la possibilità di implementare una procedura di lemmatizzazione semiautomatica tarata sulla lingua del Verismo italiano, che potrebbe essere un significativo valore aggiunto del progetto. La ricerca di una parola dà accesso ai dati quantitativi a essa relativi e alle liste di concordanza, che prevedono contesti immediati (di lunghezza stabilita di volta in volta dal consultatore), contesti allargati all'intera pagina e da qui al facsimile della stessa, con la possibilità di sfogliare l'opera in avanti e all'indietro, in modo che la contestualizzazione della parola o del fenomeno ricercato sia potenzialmente aperta a qualunque misura si riveli necessaria (vd. Fig. 2).

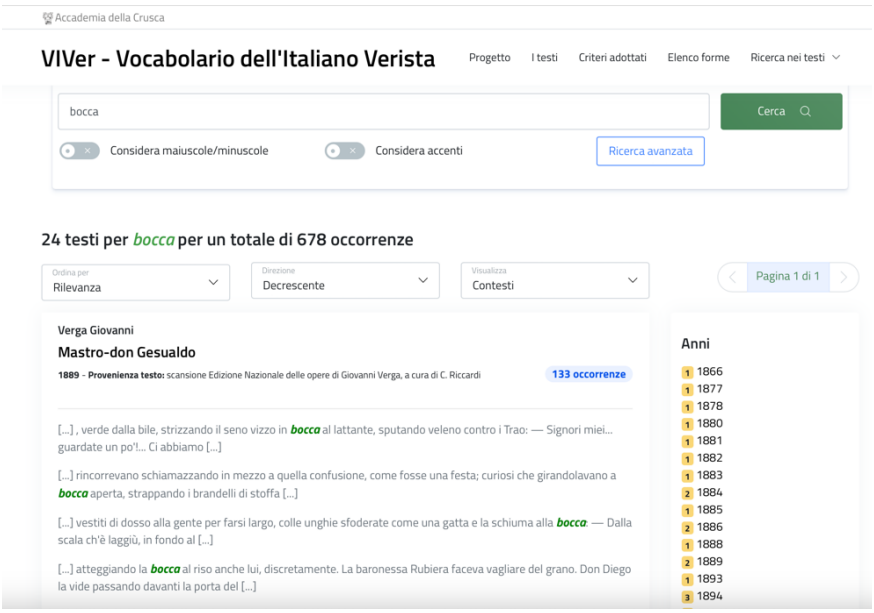


Figura 2. Schermata dei risultati della ricerca della forma bocca

A una ricerca semplice, che permette la libera interrogazione del testo a prescindere dalla griglia di marcatura, si affiancherà una ricerca avanzata, attraverso la quale sarà possibile individuare regionalismi, tecnicismi ecc., ed effettuare ricerche mirate di forme all'interno delle porzioni di testo ricondotte a una certa categoria dallo specifico marcatore. A questo approccio, funzionale per chi sappia che cosa ricercare, si affiancheranno anche ricerche guidate che accompagnino il consultatore nelle specificità della banca dati; e ricerche guidate con specifici obiettivi didattici in modo che essa possa diventare anche un importante punto di partenza per le attività di insegnamento della lingua e della letteratura nella scuola secondaria e all'università. È prevista anche una "Sala di lettura", che trasforma di fatto il corpus in una biblioteca digitale,

[3] https://ogp.me/
[4] https://developer.twitter.com/en/docs/twitter-for-websites/cards/overview/markup

dai cui "scaffali" il lettore (sia esso un addetto ai lavori o appartenente al largo pubblico) potrà estrarre uno dei testi e leggerlo contando, oltre che sul testo elettronico, anche sulla riproduzione in facsimile dell'originale.
Nella sala di lettura sono previsti tre livelli per i vari testi (che corrispondono poi alle tre fasi di lavorazione): a) accesso al PDF (accompagnato da una scheda di introduzione); b) accesso alla versione elettronica collazionata e indicizzata (e quindi già disponibile anche per la modalità ricerca), con facsimile dell'originale; c) accesso alla versione marcata. Un'apposita maschera consentirà la ricerca per autore, titolo, data, per poter così raggiungere la scheda dell'opera desiderata, che sarà disponibile in uno dei tre formati previsti. La banca dati è aperta e dinamica e pertanto i testi che di volta in volta raggiungeranno livelli diversi di lavorazione "saliranno" alla categoria successiva (dal PDF alla versione indicizzata, dalla versione indicizzata a quella marcata).
L'impostazione modulare consente di mettere a disposizione dei consultatori i materiali anche se non hanno raggiunto una completa lavorazione. Ci è sembrata infatti una scelta preferibile quella di consentire l'accesso immediato al più alto numero possibile di testi, seppure a livelli di base o intermedi (comunque dichiarati), piuttosto che attendere lo stadio finale del completamento della marcatura dell'intero corpus. La stessa filosofia è stata applicata anche agli strumenti di interrogazione, e per questo è già disponibile al pubblico il prototipo della banca dati testuale (ancora non pubblicato ufficialmente ma interrogabile all'indirizzo https://testi.progettoviver.it/). Questa versione rappresenta solamente una prima versione di quella che sarà la piattaforma di consultazione finale (vd. Fig. 3), essendo limitata nella quantità (al momento sono presenti solamente 33 su oltre 200 già individuati e in lavorazione) e nella funzionalità (i testi caricati non erano ancora marcati, quindi l'indicizzazione è stata fatta sul testo piano).

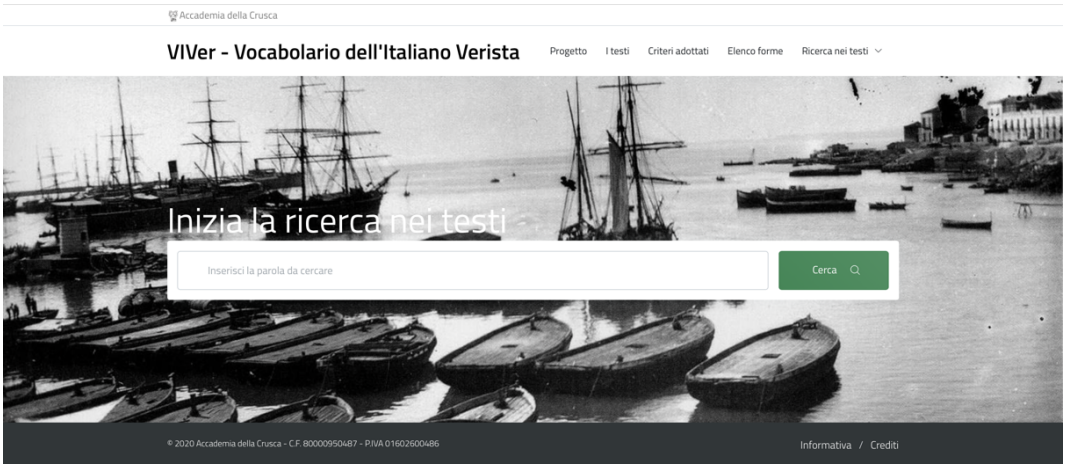


Figura 3. Pagina d'entrata del sito del VIVer

Non va dimenticato che la banca dati testuale ha come sua primaria funzione quella di fornire un corpus rappresentativo che serva di base per il dizionario. Di questo è ancora in fase di elaborazione il tracciato della voce, anche se sono chiari alcuni punti fermi che lo caratterizzeranno.
Il primo è che la struttura della voce sarà sicuramente in linea con quella dei dizionari storici (come il TLIO), a vocazione storica (come il GDLI), o insieme sincronici e diacronici come il VoDIM [5]. Saranno previsti quindi campi legati alla prima attestazione (con particolare attenzione all'ingresso in ambito verista e a eventuali specifiche accezioni), alla rete di corrispondenze, a indicazioni di tipo diatopico e diafasico, nonché agli ambiti d'uso. Sono poi allo studio campi che rendano conto delle specificità del lessico verista, anche in chiave comparativa con le letterature affini delle altre lingue europee (individuando ad esempio *cluster* multilingui di parole corrispondenti allo stesso significato che potrebbero essere fornite come possibili traducenti ideali in contesto verista in un apposito campo del dizionario).
Il secondo è la natura dinamica del dizionario (e qui il riferimento è di nuovo al VoDIM, [4, 5, 6, 7], vale a dire la possibilità di garantire una consultazione di voci a struttura variabile. A partire da una "metascheda" che raccoglie tutti i campi previsti, saranno infatti messe a punto schede mirate in funzione del consultatore (ricercatore, studente, docente, consultatore straniero, curioso, ecc.), e sarà prevista la possibilità di organizzare una voce personalizzata del dizionario scegliendo i campi di specifico interesse (si vedano come esempi i prototipi di voci dinamiche preparate per il VoDIM per quanto riguarda la cucina [4] e per l'arte [12]).
Sono infine previsti collegamenti interni, reciproci, fra il dizionario e la banca dati testuale, e collegamenti esterni con corpora e dizionari disponibili in rete.

BIBLIOGRAFIA

[1] Alfieri, Gabriella. *«La vita più spensierata del mondo». Spigolature idiolettali nel vissuto linguistico del Verga 'milanese' (1872-1891)*. Leonforte: Euno Edizioni - Fondazione Verga, 2020.

[2] Alfieri, Gabriella. *Verga*. Roma: Salerno editrice, 2016.

[3] Alfieri, Gabriella, e Giorgio Longo. «Vues et voix de l'étranger dans le Verisme italien".» (a cura di) Lumbroso Olivier. *Les Cahiers Naturalistes*, Naturalismes du monde, les voix de l'étranger, fasc. 94 (2020): 369–404.

[4] Bertini Malgarini, Patrizia, Marco Biffi, e Ugo Vignuzzi. «Dal Vocabolario storico della cucina italiana postunitaria (VoSCIP) al Vocabolario Dinamico dell'Italiano Moderno (VoDIM): riflessioni di metodo e prototipi». *Studi di Lessicografia Italiana* XXXVI (2019): 341–66.

[5] Biffi, Marco. «Progettare il corpus per il vocabolario postunitario». (a cura di) Claudio Mazzarini e Ludovica Maconi, 259–80. Firenze: Accademia della Crusca, 2016.

[6] Biffi, Marco «Strumenti informatico-linguistici per la realizzazione di un dizionario dell'italiano post-unitario». In *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data (JADT '18)*, (a cura di) Domenica Fioredistella Iezzi, Livia Celardo, e Michelangelo Misuraca, 1:99–107. Roma: Universitalia, 2019.

[7] Biffi, Marco, e Alice Ferrari. «Progettare e ideare un corpus dell'italiano nella rete: il caso del CoLIWeb». *Studi di Lessicografia Italiana* XXXVII (2020): 357–74.

[8] «I verismi regionali. Atti del Congresso Internazionale di Studi, Catania, 27-29 aprile 1992». Catania: Fondazione Verga, 1996.

[9] Lumbroso, Olivier, (a cura di). «Naturalismes du monde, les voix de l'étranger». In *Les Cahiers Naturalistes*, 94:175–318. sez. II, 2020.

[10] Luperini, Romano. *Giovanni Verga. Saggi (1976-2018)*. Roma: Carocci, 2019.

[11] Nencioni, Giovanni. «Francesco De Sanctis e la questione della lingua». In *La lingua dei Malavoglia e altri scritti*, 237–82. Napoli: Morano, 1998.

[12] Patella, Barbara. «Il Vocabolario dinamico dell'italiano moderno (VoDIM): proposta di schede lessicografiche per la lingua dell'arte». *Italiano digitale* XIII, fasc. 2 (2020): 122–70.

[13] Salucci, Giovanni. «Utilizzo del DOI (Digital Object Identifier) nei progetti di digital humanities». *DILEF. Rivista digitale del Dipartimento di Lettere e Filosofia* 2 (2022): 308–19.

[14] Salucci, Giovanni. «Utilizzo del DOI (Digital Object Identifier) per la diffusione di progetti lessicografici digitali». *DILEF. Rivista digitale del Dipartimento di Lettere e Filosofia* 3 (2023): 275–92.

[15] Valussi, Pacifico. *Caratteri della civiltà novella in Italia*. Udine: Gambierasi, 1868.

L'informatizzazione del GDLI: risultati, prospettive, sfide future

Eva Sassolini[1], Sebastiana Cucurullo[2], Marco Biffi[3]
[1]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - eva.sassolini@ilc.cnr.it;
[2]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - sebastiana.cucurullo@ilc.cnr.it
[3]Università degli Studi di Firenze e Accademia della Crusca, Italia - marco.biffi@unifi.it

ABSTRACT[1]

In questo articolo vogliamo porre l'attenzione sulle problematiche relative alla costruzione di una banca dati digitale in un contesto in cui i dati non hanno una rigida strutturazione e sono affetti da errori di tipo ortografico e sulla costruzione di un sistema software per la loro consultazione online, condividendo con la comunità scientifica le attività, i metodi, i risultati intermedi e le prospettive che caratterizzano il progetto di informatizzazione del *Grande Dizionario della Lingua Italiana di Salvatore Battaglia* (GDLI). Il progetto, nato dalla collaborazione tra l'Accademia della Crusca e l'Istituto di Linguistica Computazionale "A. Zampolli" del CNR (CNR-ILC), si pone l'obiettivo di realizzare una banca dati interrogabile online con funzionalità di ricerca avanzate. Attualmente la disponibilità di dizionari digitali in Internet registra l'accesso di una vasta platea di utenti e anche il GDLI, da quando è accessibile online nella versione non strutturata, è molto consultato. Il progetto vuole offrire un punto di accesso più articolato e utile agli studi di linguisti e storici della lingua. L'analisi condotta sui dati e sulle potenzialità di sfruttamento delle informazioni ivi contenute ha orientato lo sviluppo verso soluzioni e strategie implementative capaci di controllare e gestire gli errori e costruire un modello ibrido di rappresentazione dei dati, scomposto in più componenti fisici collegati ma autonomi; una risorsa multidimensionale in grado di produrre prospettive di analisi e di consultazione diversificate del dizionario. L'aspetto realizzativo della banca dati ha condizionato la scelta delle funzionalità del sistema di interrogazione, che sono diventate necessariamente dedicate, con specifiche di realizzazione distinte per le diverse dimensioni (componenti) da indagare. In questo scenario l'implementazione software ha seguito un approccio sperimentale, un punto di vista che permettesse di procedere a stadi di avanzamento progressivo e ad una verifica costante delle scelte adottate. Con queste premesse la conclusione del progetto è un obiettivo di medio lungo termine ma che va nella direzione di favorire la valorizzazione di questa importante risorsa lessicografica.

PAROLE CHIAVE

Knowledge Representation; Cultural Heritage; Information Extraction; Information Rietrieval; Digital Humanities.

1. INTRODUZIONE

Il *Grande Dizionario della Lingua Italiana* (GDLI) è un fondamentale strumento lessicografico dell'italiano in diacronia: in esso è raccolto l'intero patrimonio lessicale italiano dai primi documenti a oggi, e le sue voci riportano informazioni sull'evoluzione dei significanti e dei significati corredate da una ricchissima serie di esempi che documentano l'uso linguistico nel corso dei secoli. Si tratta di 21 volumi, con 2 supplementi e un volume dedicato allo scioglimento delle abbreviature usate per le fonti, per un totale di circa 25.000 pagine, su tre fitte colonne. Ormai da decenni sono noti i vantaggi non soltanto di dizionari concepiti e realizzati in partenza come elettronici (*primari*), ma anche di quelli *secondari* ottenuti trasformando in elettronici dizionari concepiti e realizzati come cartacei. In questo secondo caso vantaggi enormi si ottengono già semplicemente con il passaggio a un testo elettronico interrogabile come testo libero: gli strumenti della linguistica computazionale consentono infatti di dare al dizionario un valore aggiunto, che è quello di una ricerca indicizzata anche delle forme presenti nel testo e non a lemma, particolarmente utile per studiare il lessico contenuto in quei dizionari che, in particolar modo nel passato, erano non circolari per impostazione (come il *Vocabolario degli Accademici della Crusca*, uscito nella prima edizione del 1612 per descrivere parole del fiorentino del Trecento, ma ricorrendo anche al lessico della lingua contemporanea dei redattori). Ma i vantaggi si moltiplicano nel momento in cui si restituisce al dizionario la sua vera natura di database, informatizzando anche i campi, più o meno espliciti, che sottostanno allo schema di ogni voce lessicografica, consentendo un accesso indicizzato alle informazioni (categorie grammaticali, datazioni, ambiti d'uso ecc.). Il recupero della struttura potenzia in particolar modo strumenti come il GDLI, corredato di una fitta serie di esempi, tale da costituire di per sé un vero e proprio corpus rappresentativo dell'italiano in diacronia una volta che i campi che contengono gli esempi possono essere identificati e gestiti da procedure informatiche.

[1] Nel quadro di un'elaborazione comune, Marco Biffi, come responsabile del progetto per l'Accademia della Crusca, ha curato la redazione del § 1, Eva Sassolini, come responsabile del progetto per CNR-ILC, dei §§ 2 e 4; Sebastiana Cucurullo, come responsabile dello sviluppo della banca dati, del § 3.

Dal 9 maggio 2019 i volumi del GDLI sono consultabili in rete dagli "Scaffali digitali" del sito dell'Accademia della Crusca, o direttamente all'indirizzo <http://www.gdli.it/>. L'operazione è stata possibile dopo che, il 2 settembre 2017, l'Accademia della Crusca e la casa editrice UTET avevano firmato l'accordo che rendeva possibile la digitalizzazione. Questa versione è in realtà un prototipo, come si dichiara espressamente nella pagina d'entrata: si ferma al primo livello di informatizzazione (quello del testo) e il testo elettronico non è stato sottoposto a controllo. Quella dell'Accademia è stata una scelta coraggiosa, che in effetti è andata incontro a qualche critica nell'ambito dell'umanistica digitale dove non tutti hanno apprezzato la provvisorietà del testo non collazionato, rimasto allo stadio raggiunto con il semplice riconoscimento OCR (*Optical Character Recognition*). Il formato risultante, a cui sono seguiti pochissimi interventi (si sottolinea l'importanza di quelli volti a ricostruire l'unità delle parole sillabate, che altrimenti sarebbero rimaste nascoste allo scandaglio informatico), presenta sia errori ortografici che la completa mancata decodifica del testo in greco, frequente soprattutto nelle etimologie. In tale operazione ha prevalso l'intento di mettere subito a disposizione degli studiosi uno strumento così prezioso e importante, dichiarandone i limiti, piuttosto che l'attesa di anni prima di poter disporre di una versione in parte migliorata. La debolezza della versione provvisoria va del resto ridimensionata: la restituzione dei risultati è buona, anche perché l'impatto della ricerca etimologica di parole greche è basso nella consultazione di uno strumento come questo, e soprattutto perché comunque ogni ricerca approda non soltanto alla trascrizione in caratteri dell'intera pagina del dizionario, ma anche alla sua riproduzione in facsimile (immagine), e pertanto priva di errori. Quello che può succedere è che si perda l'occorrenza di qualche parola, ma una volta approdati al testo, questo può essere consultato nel pieno delle sue potenzialità, scorrendo le pagine, ingrandendole in modo da poter leggere agevolmente le fitte colonne impresse con caratteri molto piccoli spesso di faticosa lettura nella versione cartacea [3].
In contemporanea con la pubblicazione in rete del prototipo, l'Accademia della Crusca, in collaborazione con CNR-ILC, ha avviato un progetto articolato per l'informatizzazione della struttura, partendo da una prima fase in cui ci si è concentrati sull'individuazione e la marcatura delle entrate lessicali, ma prevedendo fin da subito una seconda fase in cui al campo "Lemma" si sarebbero aggiunti i campi "Definizione", "Esempio", "Nota etimologica". Già dal suo primo avvio il progetto è stato concepito come piano di lavoro a traguardi progressivi. Migrare i dati da formato digitale non standard a rappresentazione strutturata delle entrate lessicali si è dimostrata da subito un'operazione piena d'insidie e dai molteplici risvolti, sia per le condizioni del formato digitale dell'input, sia per la complessità insita nei dati originali. Il primo importante obiettivo del progetto è stato il formato standard di rappresentazione digitale, che per questa importante risorsa mancava: la conversione dei dati in formato XML/TEI codificato secondo standard consolidati pensati per i dizionari [13]. L'intero dizionario è ad oggi disponibile in una versione semplificata dell'articolazione interna della voce [2, 4], ovvero organizzato per ogni entrata nelle quattro macro-aree/dimensioni sopra indicate: lemma; informazione semantica e grammaticale (che include la definizione e la categoria grammaticale, le marche d'uso, ecc.); esempi; etimologia. La risorsa TEI strutturata in questo modo rappresenta oggi il punto di partenza di ogni successivo sviluppo ed elaborazione sul dizionario digitale.

2. LA STRUTTURAZIONE DEI DATI

Un formato più articolato, che individui tutti gli elementi rilevanti ivi contenuti, rimane un obiettivo del progetto ma in una prospettiva di più lungo termine perché necessita di ulteriori raffinamenti nelle procedure di estrazione, in larga misura per le difficoltà di gestione degli errori prodotti nella fase di acquisizione, ma non solo. Infatti, il GDLI è un'opera monumentale in cui le informazioni sono pensate e organizzate per l'utente umano, che è in grado di comprendere la tipologia dei contenuti anche in assenza per essi di una sistematica collocazione e/o di elementi specifici di formato. Se si analizza la struttura dell'entrata si osserva come posizione, stile e sequenza delle informazioni che seguono il lemma non sia rigorosa. Un esempio di questo comportamento si riscontra quando si tenta di isolare la categoria grammaticale perché, oltre a non avere una posizione fissa, a volte non è identificata da una sigla univoca; per esempio, sono equivalenti le espressioni "escl." e/o "esclam."; "cong." e/o "congiunz.". Comportamento che si ripropone anche nelle forme composte, come avviene nel caso di "cong. finale" e/o "cong. fin.". Oppure facendo seguire alla categoria grammaticale spiegazioni più o meno articolate, come per "intr. più spesso con la particella pronom." e "intr. con la particella pronom.". In altri casi si fa ricorso al suo completo scioglimento in frase, in altri ancora se ne omette ogni riferimento. Questa difformità che per un umano è facilmente disambiguabile dal contesto, per un sistema software automatico diventa difficile da gestire perché mancano le regole di riconoscimento, valide per tutte le occorrenze, in grado di guidare la procedura automatica. Anche nel caso di utilizzo di sistemi di apprendimento automatico, manca una casistica di esempi onnicomprensiva per l'addestramento [6]. La peculiare natura e formato dei contenuti ha reso quindi infruttuoso utilizzare strumenti allo stato dell'arte per l'estrazione della struttura del dizionario e ha imposto un processo iterativo, che si raffina/perfeziona passo dopo passo.
Mentre nelle fasi iniziali del progetto il gruppo di lavoro si è impegnato nella ricerca di soluzioni e strategie che mirassero a convertire i contenuti testuali in dati digitali strutturati, confidando di poter trascurare gran parte degli errori ortografici

derivanti da OCR, le progressive sperimentazioni hanno mostrato come in realtà una percentuale di essi incidesse sul processo di estrazione e ne condizionasse i risultati [10, 11]. La consapevolezza di una situazione di lavoro complessa non ha tuttavia impedito di proseguire il progetto di informatizzazione, ma ha imposto un'analisi approfondita della possibilità di rappresentare il GDLI strutturato esclusivamente attraverso un Database (DB) relazionale, più o meno articolato, come tipicamente accade per dizionari standard. Nel caso del GDLI, infatti, questa operazione si innesta in una struttura delle voci che difetta di una rappresentazione sistematica dei vari fenomeni. La modellazione dei dati, propedeutica alla progettazione fisica del DB, ha dovuto fare i conti con problematiche legate alla gestione annidata di alcune porzioni di informazioni. Per esempio, i sottolemmi, presenti in gran numero nel dizionario, non sono individuabili automaticamente anche quando non presentano errori. La loro collocazione non è segnalata da elementi strutturali o di formato che possano identificarli univocamente e spesso risultano distribuiti nelle definizioni più interne alla voce. In questo contesto è arduo ritenere che le rigide strutture del DB (che per sua natura richiede informazioni puntuali e identificabili in modo univoco) possano comprendere tutte le informazioni che la ricca struttura delle voci rivela. Il lungo lavoro di analisi dei dati e lo studio di esperienze simili [9, 5, 7, 12], anche orientate al recupero di importati risorse storiche e culturali, ha imposto uno sforzo di astrazione e suggerito di pensare il GDLI come un insieme scomponibile di conoscenza, dove ogni parte/dimensione possa essere gestita con strategie e approcci diversi. Le quattro aree già individuate nel formato TEI sono state indagate, trasversalmente lungo tutto il dizionario, al fine di comprendere quanto dei rispettivi contenuti fosse possibile scomporre ulteriormente con l'obiettivo di inserire nel DB informazioni utili a gestire più efficacemente le funzionalità di ricerca online.

Ogni macro-area ha caratteristiche peculiari e fa emergere sfide di tipo diverso. Nel campo lemma esistono diverse problematiche, in primo luogo legate alla presenza nelle entrate dell'indicazione dell'apertura/chiusura delle "e" e "o" toniche, unita a quella della natura sorda/sonora delle affricate e sibilanti. Tale caratteristica impone la predisposizione nel DB della relativa forma normalizzata (senza diacritici), che verrà utilizzata dagli utenti nella *form* di ricerca online. In second'ordine, ma non meno importante, esiste l'oggettiva difficoltà di individuare le forme alternative del lemma, i sottolemmi, le entrate multiple (di più lemmi), i rimandi. Nel campo definizione sono contenute importanti informazioni che tuttavia, non distribuendosi in sequenze fisse, è arduo marcare singolarmente: ad esempio, oltre alla categoria grammaticale, è utile individuare le marche d'uso, ma anche distinguere le definizioni principali da quelle secondarie. Per quanto riguarda l'area etimologica è in corso una campagna di correzione manuale: qui, infatti, si è concentrato il maggior numero di errori prodotti dal sistema di OCR. Anche il campo degli esempi è caratterizzato da specificità che non sono meno insidiose, soprattutto per quello che attiene all'identificazione e isolamento delle stringhe bibliografiche. Per queste ultime è stato condotto un intervento di normalizzazione/correzione manuale delle diverse forme affette da errori ortografici prodotte dal sistema di OCR[2].

3. LA BANCA DATI

La consultazione in Internet per campi rappresenta l'obiettivo primario da raggiungere, ma la scelta di una struttura composita per la banca dati ha obbligato ad elaborare strategie in grado di organizzare le informazioni che vi sono contenute. La risorsa ad oggi comprende un database relazionale contenente per ogni voce del dizionario i campi: lemma, categoria grammaticale (e informazioni grammaticali aggiuntive), definizione (distinta in principale e secondarie), etimologia. La macro-area esempi è confluita invece in un corpus testuale, agganciato al DB al livello di entrata, in cui si sono distinte le stringhe bibliografiche dalla citazione relativa. Al fine di offrire una navigazione efficiente dei dati, sono state inoltre create risorse trasversali ai campi ed esclusivamente finalizzate all'interrogazione del dizionario, elaborazioni condotte preliminarmente sul DB in grado di sintetizzare le informazioni rilevanti di ogni entrata e renderle disponibili a richiesta (per esempio: numero di definizioni; numero di esempi, ecc.). Non si sono trovati esempi di approcci simili descritti in letteratura; esistono sperimentazioni ibride su piccoli dati ma non su opere così voluminose e pertanto, i tempi di progettazione della banca dati e del relativo sistema di interrogazione hanno richiesto un impegno maggiore del previsto. L'analisi dei dati inseriti automaticamente nel DB ha mostrato criticità su alcune parti della voce e in modo specifico nell'intorno destro del lemma, dove il riconoscimento dei fenomeni presenta incongruenze: l'uso concomitante di parentesi tonde, corsivo e punteggiatura ha spesso falsato la corretta acquisizione da parte del sistema di OCR con conseguenti errori nel formato TEI da cui provengono i dati del DB. Poiché i confini di voce sono certificati da una correzione manuale, gli errori relativi ai lemmi con molte definizioni e tanti esempi sono sicuramente quelli di maggiore impatto; in queste ampie sezioni del dizionario (alcuni lemmi si distribuiscono in un considerevole numero di pagine) una non corretta acquisizione del testo può pregiudicare l'affidabilità della segmentazione interna dell'intera entrata. Per questa categoria di entrate si è

[2] Il lavoro è stato condotto in un progetto di collaborazione tra DILEF e Accademia della Crusca sotto la supervisione di CNR-ILC ad opera di Elena Pepponi.

deciso di produrre sintesi/viste grafiche preconfezionate, da recuperare/collegare in fase di interrogazione. Una risorsa così complessa come il GDLI richiede di predisporre meccanismi che migliorino l'efficienza dell'interrogazione e, dove possibile, evitare di demandare all'esecuzione a *runtime* la produzione di onerose elaborazioni computazionali.
Per testare e isolare questi errori per alcuni volumi[3] è stata prodotta una versione GDLI codificata e indicizzata secondo specifiche DBT-like [8], ma rispettando la stessa strutturazione in quattro macro-aree. L'uso del motore di analisi testuale DBT (Data Base Testuale) ha avuto il duplice scopo di testare la strutturazione realizzata, ovvero se fosse in grado di rispondere a funzionalità di ricerca avanzate, e di analizzare la qualità del mapping in XML prodotto. Attraverso le funzionalità di ricerca DBT è infatti possibile individuare in modo puntuale gli errori che permangono dopo la conversione in TEI sui quali, in caso, è necessario avviare campagne di correzione manuale mirata (vd. Fig. 1).

```
Riga   550:<quote> Quello che non le piaceva la contrariava come chi rimanga deluso da una cosa aspettata e sognata. 2. Sm. Ant. Attesa.</quote>
Riga  1118:<quote> Il Bandito ammantellato di turchino... sbuca fuori brandendo due pistole. 2. Figur. Nascosto, celato, protetto.</quote>
Riga  1974:<quote> La pietra dei gradini, ...gareggia d'asprezza con la scorza dei platani venerandi. - Figur.</quote>
Riga  5750:<quote> Era sommamente emaciato... ed atrofico in modo, che pareva uno scheletro coperto d'arida pelle. 2. Figur.</quote>
Riga  7065:<quote> Stette per ispazio d'un'ora anzi che fosse legato' Prep. di luogo. Ant. Davanti a, alla presenza di.</quote>
Riga  8338:<quote> Avvolsi con cura l'orologio in un foglio di carta velina. - Figur.</quote>
Riga  9933:<quote> Per prolungare il sorso, contenevano il respiro finché non si sentivan morire d'ambascia. - Medie. Ant. Asma.</quote>
Riga 10158:<quote> Luciferi ammorzati, Esperi ardenti. - Figur.</quote>
Riga 11486:<quote> Era un uomo giovane, molto grande e aitante nella persona, con un viso roseo, fresco e virile. 2. Sm. Disus. Aiutante.</quote>
Riga 17480:<quote> Fortuna che il tuo sole è stato onesto e rispetta una povera bionda che non ha più vestiti. Figur.</quote>
```

Figura 1. Nel formato TEI del GDLI individuazione di alcune definizioni rimaste nel campo esempi

Per quanto concerne il campo esempi la trasformazione in corpus testuale ha mostrato da subito grandi potenzialità per lo studio della lingua italiana. I contenuti sono infatti uno dei più grandi e ricchi insiemi di testi altamente rappresentativi della nostra lingua, dove DBT rappresenta certamente lo strumento di studio più adatto e flessibile.
Il lavoro di scomposizione del dizionario in parti, oltre ad individuare e risolvere i problemi annidati nei campi, ha prodotto una proficua parallelizzazione delle fasi di lavoro, condizione auspicabile con dati di tali dimensioni. Si tenga conto che ogni volume codificato in XML ha una dimensione cha varia tra i 41 e i 43 Megabyte (mb), che proiettata sui 21 volumi rende i dati complessivi di poco inferiori al Gigabyte (gb). Con queste dimensioni anche lavorare su porzioni limitate ma distribuite sull'intero dizionario rimane un'impresa impegnativa, ma consente di concentrare gli interventi più puntualmente e operare scelte diverse a seconda dei casi.
Attualmente il primo prototipo della banca dati è pronto e vi stiamo riversando tutti i 21 volumi dell'opera con un processo iterativo che mira a testare l'omogeneità, la correttezza e la dimensione dei dati. Occorre considerare che non tutti i volumi hanno ottenuto lo stesso livello di resa dal sistema di OCR. Esistono non trascurabili porzioni del dizionario in cui la percentuale degli errori di riconoscimento ha avuto un impatto maggiore, forse per lo stato di conservazione del singolo volume (colore della carta, usura del testo delle pagine, minimi effetti di curvatura delle immagini nelle porzioni centrali del volume), oppure per errate interpretazioni di peculiari e concomitanti difformità grafiche nelle voci. Per queste aree, comunque limitate, sarà necessario correggere manualmente le porzioni di testo. Qualche osservazione sui numeri del GDLI può essere già fatta visto lo stato di avanzamento dei lavori sulla banca dati. Le citazioni per lemma sono in media 7/8 ma in una distribuzione variabile sia per lemma che tra i volumi; la percentuale dei lemmi che non presentano citazioni è all'incirca del 30%, considerando tra questi anche i lemmi di rinvio, calcolo che presenta notevoli variabilità sui singoli volumi. Le entrate con almeno una citazione sono mediamente il 25%. Per quanto riguarda le definizioni per lemma una stima della media sull'intero GDLI è poco significativa perché dipende dai lemmi contenuti nei volumi. Per esempio, nel volume V sono presenti lemmi come "essere" e "fare" che hanno rispettivamente: ~190 definizioni corredate da poco meno di 1.000 citazioni il primo, oltre 350 definizioni con oltre 2.250 citazioni il secondo. Anche la percentuale dei lemmi con più di 30 definizioni mostra variazioni apprezzabili tra i volumi ma mediamente si attesta sotto il 10%.

4. L'APPROCCIO PER LA CONSULTAZIONE ONLINE

Pur avendo in mente cosa è stato fatto per altri grandi dizionari a livello internazionale, per esempio nel *Dizionario Storico della Lingua Spagnola (DHLS)*[4][5]; per le *Trésor de la langue française* (TLFi)[5] o, a livello nazionale, per il *Tesoro della Lingua Italiana delle Origini* (TLIO)[6] e per la versione elettronica in rete del *Vocabolario degli Accademici della Crusca* (*Lessicografia della Crusca in rete*)[7][1], non è stato possibile affidarsi a modelli ed esperienze consolidati. Con la consapevolezza che la completa potenzialità di un dizionario digitale risiede nella disponibilità di contenuti lessicografici

[3] La versione DBT-like comprende solo I, II volume e porzioni ragionate dei volumi dove il DB ha prodotto incongruenze.
[4] https://www.rae.es/dhle/
[5] http://atilf.atilf.fr/tlfi/
[6] http://tlio.ovi.cnr.it/TLIO/
[7] http://lessicografia.it

strutturati, diventa quanto mai determinante migliorare la qualità della codifica di questi contenuti lessicografici poiché questa condiziona la tipologia di interrogazioni possibili. Inoltre la gestione di una risorsa composita richiede di attuare una parallelizzazione delle attività, ovvero poter lavorare al DB e al corpus testuale allo stesso tempo. Di conseguenza, dal punto di vista del sistema di interrogazione, questo si traduce nel testare funzionalità di ricerca/consultazione specifici a seconda della porzione/dimensione dei dati interrogata. Le prove e le sperimentazioni sono state inizialmente condotte sui primi volumi, intervenendo con correzioni manuali nei casi in cui gli errori inficiavano i risultati del testing. Lo scenario costituito da risorse integrate non è comunque il solo problema da gestire. Come affermato in precedenza una strutturazione 'fine' in campi rappresenta una forzatura nel GDLI e questo impone di individuare soluzioni che ne tengano conto. Per le quattro dimensioni le soluzioni operative scelte sono state diverse ma permangono questioni che vanno approfondite. Per quanto riguarda il corpus degli esempi resta da valutare la gestione di elementi duplicati, ovvero frasi/citazioni che vengono riproposte nel dizionario come esemplificazione di parole diverse. Norme di efficienza suggerirebbero di evitare le duplicazioni mantenendo solo un rimando ai diversi lemmi che riferiscono la stessa citazione. In realtà però non si ha certezza dell'effettiva equivalenza di due frasi, visto che la presenza di errori ortografici potrebbe falsare sia il testo della citazione che la stringa bibliografica. Informazioni certe sulla dimensione di questo fenomeno non sono disponibili e quindi eliminare i duplicati potrebbe avere un impatto trascurabile sulle dimensioni del corpus e risultare uno sforzo inutile. Anche relativamente alle informazioni gestite con DB si possono fare alcune importanti considerazioni. È infatti possibile specializzare le procedure di estrazione all'interno dei singoli campi con regole *ad hoc*, al fine di offrire all'interrogazione online informazioni più strutturate, ma occorre capire se è sempre opportuno e con quali risultati. Un esempio concreto del problema è mostrato in Figura 2, dove si vede chiaramente come l'isolamento della categoria grammaticale (caso 1 nella figura) produce un effetto negativo sulla definizione, che ne risulta mutilata e poco comprensibile.

Figura 2. Nel formato TEI esempio di lemma con estrazione della categoria grammaticale dalla definizione

La scelta migliore è invece mantenere la definizione così come trovata nel dizionario ed operare l'estrazione della categoria grammaticale solo a fini di filtro sui dati. Alla luce di queste considerazioni, se si vuole conservare l'allineamento della banca dati con il formato TEI, occorre modificarlo, per esempio inserendo l'attributo "opt" (caso 2 in Fig. 2) per mantenere traccia di questa 'duplicazione'. La decisione implica però un'elaborazione ulteriore dei dati che suddivida le informazioni in modo diverso da come si trovano nel GDLI. In aggiunta è necessario valutare come utilizzare il filtro quando la categoria non è univocamente riconosciuta.

Sempre con l'intento di favorire la comprensione dell'utente, particolare attenzione è stata riservata alla predisposizione di funzionalità di supporto alla consultazione in grado di guidare l'utente nella navigazione della banca dati, spiegando dove necessario il formato e l'eventuale ambiguità delle risposte.

5. CONCLUSIONI E PROSPETTIVE

Grazie al supporto di un gruppo di lavoro multidisciplinare le scelte progettuali sono nate dal confronto tra le esigenze degli utenti e le specifiche software dei sistemi, condizione determinante per una valutazione obiettiva di quanto implementato. Le analisi e i controlli che si stanno conducendo sul sistema di consultazione del dizionario mirano a dimostrare che le diverse componenti possono essere interrogate in modo integrato ed efficiente. Il testing è stato quindi articolato in livelli, in modo tale che gli eventuali problemi possano in caso trovare soluzioni mirate. Tuttavia l'impatto di una interrogazione 'federata' sull'intero GDLI non è ancora valutabile in tutta la sua complessità poiché le campagne di

correzione manuale delle porzioni del testo individuate come problematiche sono tuttora in corso. Inoltre, l'analisi complessiva dell'interazione tra sistema d'interrogazione e banca dati non è terminata: resta da capire quale impatto potrà avere la gestione di risorse di natura diversa sulla manutenzione della banca dati a medio lungo termine.

Le prospettive future di sviluppo del progetto intendono da un lato proseguire con il raffinamento della strutturazione[8], dall'altro selezionare le migliori soluzioni operative, avvalendosi dell'esperienza nello studio e realizzazione di strumenti di analisi testuale di CNR-ILC. Mettere a confronto strategie e strumenti ha permesso di affrontare la sfida con un approccio incrementale [12], ma anche di mantenere un punto di vista critico su quanto sviluppato, al momento optando per questo approccio ma essendo pronti a modificarlo in futuro. Infatti l'obiettivo di rendere interrogabile online il GDLI strutturato rimane primario, ma è considerato altrettanto importante diffonderne l'uso, non solo tra gli studiosi e storici della lingua, ma anche ad un pubblico più vasto e infine renderlo interoperabile con risorse simili.

L'informatizzazione della struttura di un dizionario così monumentale, fino a qualche anno fa reperibile solo nelle biblioteche più grandi e ad oggi consultabile solo per forma, così che possa allinearsi ad altri dizionari simili disponibili in rete, con funzionalità avanzate ma nel rispetto della peculiare natura storica, è un obiettivo ambizioso; ma il progetto *in itinere* raccoglie oggi i primi frutti di un grande lavoro corale e interdisciplinare.

BIBLIOGRAFIA

[1] Biffi, Marco. "Strumenti informatico-linguistici per la realizzazione di un dizionario dell'italiano post-unitario." In *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data (JADT '18)*, a cura di Domenica Fioredistella Iezzi, Livia Celardo, e Michelangelo Misuraca, 1:99-107. Roma: Universitalia, 2019.

[2] Biffi, Marco, Francesca De Blasi, Manuel Favaro, Elisa Guadagnini, Simonetta Montemagni, e Eva Sassolini. "Parole in rete / reti di parole. Possibili impieghi didattici dei grandi vocabolari storici digitalizzati." *ITALIANO A SCUOLA* 4, no. 1 (2023): 143-188.

[3] Biffi, Marco, e Elisa Guadagnini. "Le citazioni riconducono il dizionario nell'ambito della letteratura e della vita»: Un primo sguardo d'insieme sui citati del «GDLI»." *Studi Di Lessicografia Italiana* XXXIX (2022): 351-386.

[4] Biffi, Marco, Elisa Guadagnini, Simonetta Montemagni, e Eva Sassolini. "Il lemmario del «GDLI»: dati quantitativi e prime osservazioni." *Studi Di Lessicografia Italiana* XL (2023): 331-351.

[5] Fuertes-Olivera, Pedro A., e Sven Tarp. Theory and Practice of Specialised Online Dictionaries: Lexicography Versus. Germania: Gruyter, 2014.

[6] Khemaklhem, Mohamed, Luca Foppiano, e Laurent Romary. "Automatic Extraction of TEI Structures in Digitized Lexical Resources Using Conditional Random Fields." In *Electronic Lexicography in the 21st Century: Proceeedings of ELex 2017*, edited by Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, and Vít Baisa, 598-613. Leiden - Brno: Lexical Computing, 2017.

[7] Monteleone, Mario. *Lessicografia e dizionari elettronici. dagli usi linguistici alle basi di dati lessicali.* Napoli: Fiorentino & New Technology, 2003.

[8] Picchi, Eugenio. "D.B.T.: A Textual Data Base System," II., Computational Lexicology and Lexicography, Special issue dedicated to Bernard Quemada:77-105. Pisa: Linguistica Computazionale, 1991.

[9] Salvatori, Enrica, Roberto Rosselli Del Turco, Chiara Alzetta, Chiara Di Pietro, Chiara Mannari, e Alessio Miaschi. "Il Codice Pelavicino tra edizione digitale e Public History." *Umanistica Digitale*, October 1, 2017, No 1 (2017). https://doi.org/10.6092/ISSN.2532-8816/7232.

[10] Sassolini, Eva, e Marco Biffi. "Strategie e metodi per il recupero di dizionari storici," 2020, 235-239. https://doi.org/10.6092/unibo/amsacta/6316.

[11] Sassolini, Eva, Marco Biffi, Francesca De Blasi, Elisa Guadagnini, e Simonetta Montemagni. "La digitalizzazione del GDLI: un approccio linguistico per la corretta acquisizione del testo?" In *AIUCD 2021: DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale. Raccolta degli abstract estesi della 10a conferenza nazionale*, a cura di Angelo Mario Del Grosso, Federico Boschetti, e Enrica Salvatori, 159-166. Pisa, 2021. https://aiucd2021.labcd.unipi.it/book-of-abstracts/.

[12] Sassolini, Eva, Sebastiana Cucurullo, e Alessandra Cinini. "I corpora digitali: dall'obsolescenza tecnologica, alla salvaguardia e alla condivisione." In *Conferenza GARR Selected Papers, Associazione Consortium GARR, Roma, maggio 2017*, 31-35. Roma: Consortium GARR, 2017.

[13] Sassolini, Eva, Anas Fahad Khan, Marco Biffi, Monica Monachini, e Simonetta Montemagni. "Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study." In *Electronic Lexicography in the 21st Century. Proceedings of the ELex 2019 Conference. 1-3 October 2019, Sintra, Portugal*, 603-621. Lexical Computing, 2019.

[8] È allo studio un metodo per collocare temporalmente le citazioni grazie al volume dei citati, pur prevedendo solo la correzione dell'autore nelle stringhe bibliografiche.

La digitalizzazione del dizionario latino Lana 1978

Francesca Michelone
Università del Piemonte Orientale, Italia - francesca.michelone@uniupo.it

ABSTRACT

Il testo ha l'obiettivo di mostrare attraverso un caso studio una possibile metodologia applicabile alla creazione di un dizionario digitale in accesso aperto a partire da un testo cartaceo già esistente. In particolare, è esposta l'operazione di digitalizzazione di una parte del dizionario latino pubblicato da Italo Lana nel 1978. La sezione digitalizzata corrisponde all'ambito semantico della natura nel mondo antico. L'argomento è approfondito tramite lo studio della letteratura e il risultato delle riflessioni è visibile nella codifica del dizionario stesso.

L'approccio al testo nasce dallo studio delle migliori pratiche attualmente diffuse nella lessicografia digitale e si propone di essere allineato alla produzione di dati FAIR. Tale metodologia è elaborata nel contesto specifico della lessicografia latina, ma i risultati ottenuti possono essere un esempio replicabile per chiunque affronti questo genere di operazione.

PAROLE CHIAVE

Dizionari digitali; TEI Lex-0; Dizionario bilingue latino-italiano; Natura nel mondo antico; Accesso aperto.

1. UN DIZIONARIO LATINO DIGITALE

Il presente contributo si propone di mostrare e discutere la metodologia adottata per creare un Dizionario Digitale Aperto che possa essere non solo uno strumento di consultazione ma un punto di incontro per studiosi. L'oggetto dello studio è la digitalizzazione del dizionario latino-italiano pubblicato da Italo Lana nel 1978 e, in particolare, di tutti i lemmi che afferiscono all'ambito semantico naturale[1], allo scopo di creare un dizionario digitale ad accesso aperto contenente il lessico utilizzato nella letteratura latina per descrivere l'ambiente. Si tratta di una precisazione necessaria perché ha portato a organizzare i lemmi del dizionario anche secondo categorie concettuali che creano un ulteriore livello di lettura del testo, diverso da quello alfabetico. Sono discussi di seguito i singoli passaggi con le rispettive criticità e soluzioni adottate. Ogni fase, dall'acquisizione tramite OCR alle scelte per la visualizzazione e pubblicazione, ha seguito un percorso che si ritiene possa esemplificare una metodologia concreta per la digitalizzazione di un dizionario.

Nel seguente contributo non si affronta il tema della creazione di un dizionario nativamente digitale, tipologia che aprirebbe a riflessioni e strumenti diversi. Per questo scopo sono un riferimento i tools del gruppo *Elexis* [6], un punto di partenza e confronto per la lessicografia digitale. Tra di essi è necessario menzionare *Elexifier,* lo strumento volto a trasformare dizionari già esistenti secondo gli standard di *Elexis*. Il materiale preesistente può essere in formato .pdf o .xml, ed è possibile apportare modifiche ai dati acquisiti tramite *Lexonomy,* l'applicazione dedicata appunto alla scrittura dei dizionari. Un esempio virtuoso per la creazione di un dizionario digitale a partire da un modello cartaceo è dato dal progetto sul dizionario storico portoghese MOR Digital [1]. Per la sua creazione è utilizzata la codifica specifica TEI LEX-0, il cui uso è descritto nel dettaglio da Toma Tasovac sulla piattaforma DARIAH [4] ed è anche il riferimento per *Elexis*.

Nel campo della lessicografia, focalizzando l'attenzione sulle risorse specifiche create per la lingua latina, un importante strumento è dato dal progetto ERC LiLa-Linking Latin[2] [3], conclusosi nell'estate 2023, che raggruppa e permette di ricercare in maniera avanzata più risorse lessicografiche legate al latino. Per il latino classico il dizionario bilingue di riferimento è quello edito da Lewis-Short nel 1879, la stessa edizione presente nella *Perseus Digital Library*[3] e in *Logeion*[4]. All'interno di quest'ultimo sono presenti anche dizionari dal latino al francese (Du Cange del 1887 e Gaffiot del 1934). Quelle citate sono tutte risorse ad accesso aperto, ma occorre ancora ricordare la versione digitale del *Thesaurus Linguae Latinae*[5] consultabile online in due modalità: ad accesso aperto in formato .pdf scaricandolo dal sito del Thesaurus, oppure a pagamento nel sito dell'editore De Gruyter in una versione basata su annotazione XML/TEI. Il TLL non è solo il dizionario latino più completo attualmente disponibile, perché raccoglie tutti gli usi e le costruzioni di ogni singola parola, ma nei suoi articoli ne traccia anche una sorta di "biografia", a partire dalla comparsa nella lingua e dall'etimologia, fino all'uso nel latino tardo; si tratta quindi di una risorsa online estremamente preziosa.

[1] Questo lavoro si colloca all'interno di un progetto di dottorato con borsa PON a tematica GREEN.
[2] «LiLa: Linking Latin». Università Cattolica del Sacro Cuore. https://lila-erc.eu
[3] Crane, Gregory R. «Perseus Digital Library». Tufts University. https://www.perseus.tufts.edu/hopper/
[4] «Logeion». University of Chicago. https://logeion.uchicago.edu
[5] «TLL Open Access: Thesaurus Linguae Latinae». Bayerische Akademien der Wissenschaften. https://thesaurus.badw.de/tll-digital/tll-open-access.html.

Da questa ricognizione emergono due aspetti: il primo è come gli strumenti lessicografici dedicati al latino siano un ambito vivo e in continuo sviluppo[6]; il secondo consiste nell'assenza dell'edizione di un dizionario online ad accesso aperto latino-italiano recente. Il progetto presentato in questa sede si propone proprio di colmare questo vuoto partendo da un dizionario già esistente a stampa, appunto il Lana 1978 pubblicato dall'editore Paravia che dopo essere stato a catalogo per alcuni anni tornò nelle mani del suo autore e degli eredi che hanno dato il consenso alla digitalizzazione per finalità di studio e ricerca sotto licenza Creative Commons CC BY- NC – SA.

Le fasi principali del lavoro di digitalizzazione di un dizionario, che verranno articolate nelle successive sezioni di questo contributo, sono:

- Acquisizione e correzione del testo
- Codifica
- Restituzione digitale.

2. ACQUISIZIONE E CORREZIONE DEL TESTO

Una significativa riduzione degli errori di riconoscimento è conseguita alla manipolazione delle immagini acquisite: esse sono state ingrandite al 200% con filtro bicubico e ritagliate in modo da eliminare il margine (vd. Fig. 1). Questi accorgimenti apportati in collaborazione con lo sviluppatore Massimo Ghisalberti hanno facilitato l'acquisizione del testo.

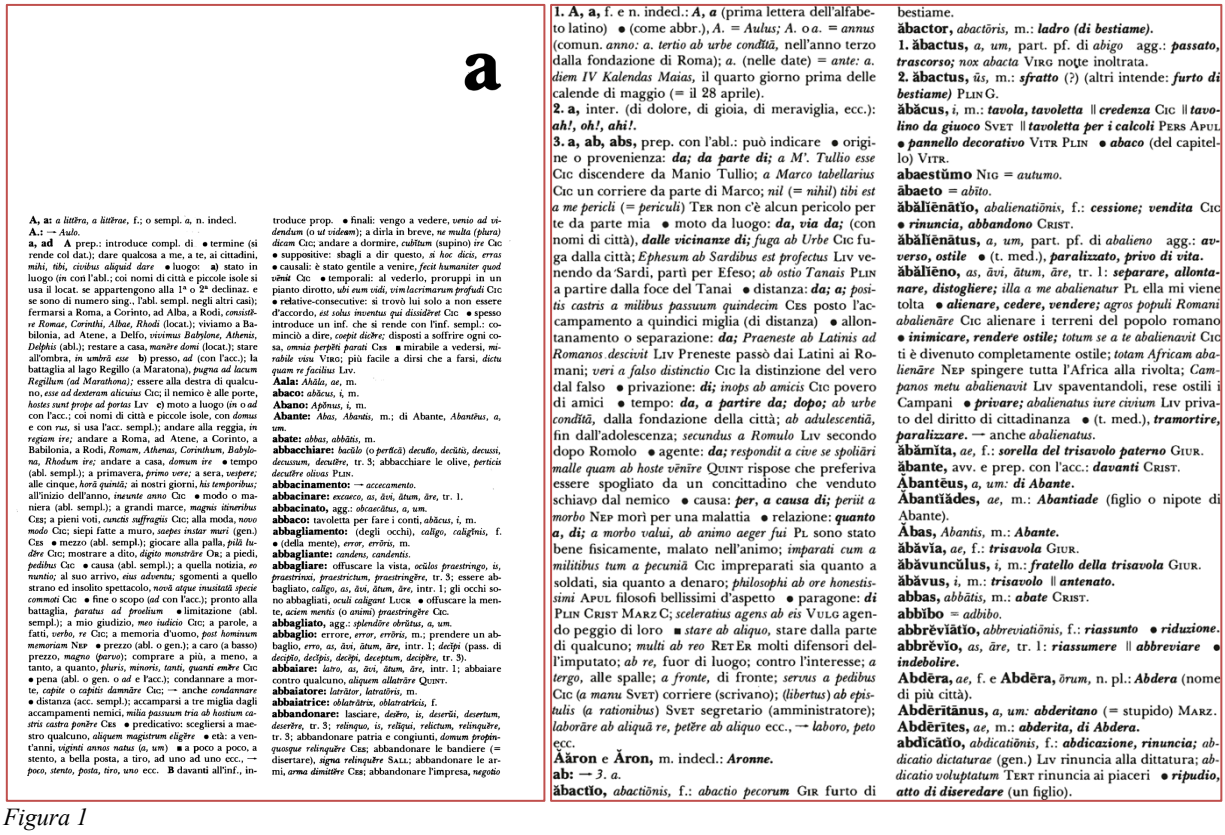


Figura 1

I programmi di OCR siano essi a pagamento o ad accesso aperto, sono numerosi. La scelta del programma dipende dal materiale di partenza: nel caso del dizionario Lana 1978 si tratta di un testo bilingue a stampa su due colonne. Ad un primo sguardo, l'acquisizione tramite OCR di un testo moderno non dovrebbe presentare complicazioni perché la stampa è di buona qualità su carta anch'essa di buona qualità. Nonostante ciò, il dizionario Lana 1978 ha richiesto la manipolazione delle immagini descritta e l'utilizzo di un programma di OCR molto configurabile: la presenza di caratteri speciali al suo interno e l'uso diffuso dei segni per indicare le quantità vocaliche creano la necessità di un *tool* specifico, come potrebbe accadere anche per dizionari che abbiano ad esempio al loro interno i caratteri dell'alfabeto fonetico.

[6] Si pensi anche al lavoro sulla metrica portato avanti da *Musisque Deoque*, Università Ca' Foscari. (https://mizar.unive.it/mqdq/public/index).

Quando il progetto ha avuto inizio nel 2022 le due alternative adatte a questo dizionario potevano essere *Transkribus* e *Tesseract*. Da un lato *Transkribus*[7] è una piattaforma nata per la trascrizione dei manoscritti di semplice utilizzo e negli ultimi anni con l'implementazione di processi di Machine Learning si prospetta sempre più avanzata e predisposta alle personalizzazioni dei modelli; la piattaforma prevede un sistema di crediti mensili gratuiti ai quali se ne possono aggiungere altri a pagamento per l'acquisizione del testo dalle immagini. Di fatto, per digitalizzare un dizionario sarebbe necessario versare un contributo per il numero di pagine e plausibilmente per una buona customizzazione del modello. Dall'altro lato *Tesseract* consente l'accesso al codice sorgente e, se si hanno le competenze, permette un livello di personalizzazione estremamente dettagliato. Per *Tesseract* sono disponibili dei file .traineddata per le diverse lingue ed è possibile modificarli secondo le esigenze. Nel caso del dizionario Lana si è optato per l'uso di *Tesseract* 5.3.4 abbinato ai dati Latin (4.1.0) e *Latincustom creati* su misura da *Rescribe* Ltd[8] [5]. Tale scelta è stata presa in conformità con le fasi di correzione e codifica cui fornisce i file .txt che sono il punto di partenza per l'elaborazione successiva che avviene tramite il *software Dicnorm*[9] prodotto dallo sviluppatore Ghisalberti. Questo programma agisce in varie fasi:

1) analisi del sorgente grezzo derivato dall'OCR;
2) generazione di un file normalizzato, applicando regole di sostituzione di caratteri o parole, individuazione degli autori in base a una tabella di sostituzioni, riunione di linee spezzate;
3) analisi del file normalizzato e produzione dell'*output*.

Pare importante sottolineare come la seconda fase, tramite un file di configurazione .yml, consenta di procedere ad una correzione per la maggior parte automatica. Se da una parte il file .traineddata addestrato da *Rescribe* ha ridotto il numero di errori di riconoscimento, il punto di svolta è stato l'applicazione del *software* per la normalizzazione e l'eliminazione degli errori ricorrenti.
Da *Dicnorm* vengono prodotti diversi file intermedi per un'ulteriore revisione manuale e i file annotati in .xml la cui codifica è descritta nel paragrafo successivo.
Siccome il primo risultato della digitalizzazione del Lana 1978 sarà un dizionario specialistico delle parole utilizzate in latino per descrivere la natura e l'ambiente, al lavoro di correzione in questo caso si è accompagnato quello di selezione dei lemmi.

3. CODIFICA

Perché il dizionario sia effettivamente digitalizzato è necessaria una codifica del testo dopo che esso sia stato acquisito e corretto. La codifica condivisa per i dizionari digitali è, come anticipato, quella nota con il nome di TEI Lex-0, una selezione pensata di XML/TEI che stabilisce uno standard per l'uso dei marcatori nei dizionari. L'elemento critico che emerge, come però già nella fase precedente, è la mole di materiale da annotare. Si tratta di un'operazione difficilmente gestibile manualmente, e si è dunque cercato un sistema che permettesse una codifica automatica del testo. Per questo scopo si sarebbe potuto utilizzare GROBID[10] "a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents" che deve essere addestrato su un campione di pagine contenenti l'annotazione richiesta in modo che essa possa essere replicata sull'intero documento. Si tratta di un tool open source che però, nonostante la ricca documentazione disponibile, necessita di una formazione specifica per essere usato in modo produttivo. Dalla collaborazione con Ghisalberti, l'informatico già citato, è nata una strategia alternativa: l'inserimento di tag e marcatori in modo semiautomatico tramite l'utilizzo di script che operano in presenza di specifici elementi del testo. Questo purtroppo non consente di riportare tutta la complessità di TEI Lex-0 nell'annotazione, ma è utile per avere almeno una prima annotazione di base da poter arricchire e specificare ulteriormente, se si ritiene, in un secondo momento.

Qui di seguito si propone un lemma campione annotato con le indicazioni base di TEI Lex-0 applicate in modo automatico con *Dicnorm*:

```
<entry lang="lat" id="L.clq2pmd50000974lazapx0flz"> <lbl ana="head">verbēnae</lbl>
<form type="taxonomy">
<lbl type="category">piante</lbl> </form>
<form type="lemma">
<orth lang="lat">verbēnae</orth> <orth lang="lat">ārum</orth>
```

[7] «Transkribus». READ-COOP. https://readcoop.eu/it/transkribus/.
[8] «Rescribe OCR». Durham University. https://rescribe.xyz/.
[9] https://gitlab.com/minimalprocedure/dicnorm
[10] Lopez, Patrice. «Kermitt2/Grobid». https://github.com/kermitt2/grobid

```
  </form>
  <gramGrp>
<gram>f.</gram>
<gram>pl.</gram>
</gramGrp>
<sense id="L.clq2pmd50000974lazapx0flz.nor">
<metamark function="normal"></metamark>
<def>ramoscelli sacri (di alloro, mirto, olivo e di altre piante che crescevano in luoghi
sacri)</def>
<note motivation="editing">
<span>ramoscelli sacri (di alloro, mirto, olivo e di altre piante che crescevano in
luoghi sacri)</span> </note>
</sense>
<sense id="L.clq2pmd500009742azapx0flz.nor">
<metamark function="normal"></metamark>
<def> piante medicinali (olivo, mirto, edera ecc.) </def> <note motivation="editing">
<span> piante medicinali (olivo, mirto, edera ecc.) </span> </note>
<cit resp="Michelone">

<quote lang="lat"> Quotiens autem medicamentum inicitur, diluendum est uel cremore
lenticulae uel aqua, in qua aut eruum aut oleae uerbenaeue decoctae sint
</quote>
<quote lang="it"> Ogni volta che venga somministrato un rimedio, esso deve essere
diluito o in una crema di lenticchie o in acqua nella quale siano stati bolliti o le vecce
o gli estratti delle piante medicinali</quote>
<bibl>
<author ref="##clq2pmd0j000970la7xccrvuv">Celso</author>
<title lang="lat"></title>
<publisher></publisher>
<date></date>
<biblScope></biblScope>
</bibl>
</cit>
</sense>
<sense id="L.clq2pmd50000974lazapx0flz.loc">
<metamark function="locuz"></metamark> </sense>
<sense id="L.clq2pmd50000974lazapx0flz.pgr">
<metamark function="pgram"></metamark> </sense>
<sense id="L.clq2pmd50000974lazapx0flz.see">
<metamark function="see"></metamark> </sense>
<sense id="L.clq2pmd50000974lazapx0flz.equ">
<metamark function="equal"></metamark> </sense>
</entry>
```

`entry` è il tag per indicare l'unità base del dizionario, il lemma. Al suo interno presenta due attributi la cui presenza è obbligatoria:

- `@xml:id` che permette il riconoscimento univoco del lemma;
- `@xml:lang` che indica invece la lingua di appartenenza, in questo caso il latino.

Il tag `form` contiene `orth` e l'attributo `@type=lemma` che servono a dare la forma ortograficamente corretta del lemma. TEI Lex-0 raccomanda l'uso del tag `form` con `@type=inflected` o `@type=paradigm` per sostantivi, aggettivi e verbi che subito dopo la forma ortograficamente corretta presentino una o più forme flesse. In una codifica automatica però a causa della varietà delle forme e delle numerose irregolarità presenti non è stato possibile avere questo dettaglio di codifica. Si è dunque optato per una versione semplificata, comunque compatibile e che distinguesse le varie forme, senza però specificarle.

Nel gruppo racchiuso dal marcatore `GramGroup` sono contenute le informazioni grammaticali. In questo punto dovrebbe essere inserita la parte del discorso (POS), ma nel dizionario latino digitalizzato essa è spesso sottintesa perché evidente,

ad esempio, dal paradigma. Sono state quindi inserite qui in modo automatico le informazioni relative al gruppo grammaticale di appartenenza sulla base del fatto che dopo la forma base del lemma, nel dizionario vengono date le informazioni grammaticali e quindi esse sono annotate automaticamente.
Un altro *tag* obbligatorio in TEI Lex-0 è `sense`, che viene utilizzato per definire il significato o dare una traduzione. Ogni `sense` ha un suo `@xml:id` univoco, generato automaticamente.
Due attributi previsti dallo *standard,* sono utilizzati in modo originale. Il primo è `@type=category` che colloca il lemma all'interno di una o più categorie semantiche, nel caso dell'esempio le piante. Il presente dizionario segue l'ontologia di Hallig-Wartburg (1963) [2] che fornisce uno schema gerarchico di concetti e un vocabolario controllato; essa è stata trasportata in OWL e funge come punto di riferimento per l'analisi di testi storici[11]. Per questo dizionario si è fatto riferimento alla sua parte dedicata all'universo e secondariamente all'uomo. I lemmi del Lana 1978 sono stati suddivisi in cinque categorie assegnate durante la fase di selezione e annotate automaticamente: cielo e atmosfera, superficie terrestre, piante, animali, essere umano. L'inserimento di queste o altre categorie in un dizionario permette una ricerca mirata e un livello di lettura ulteriore. Questo elemento, dopo aver portato nella dimensione digitale i lemmi tramite la codifica, arricchisce il dizionario aprendolo all'interoperabilità con altri strumenti del web semantico.
Il secondo attributo utilizzato è `@resp` che nelle citazioni identifica chi ha inserito e tradotto una determinata frase per esemplificare l'uso del vocabolo. Automaticamente è inserita per tutti i lemmi la responsabilità all'autore originario, Italo Lana. Tramite questo attributo è possibile identificare chi arricchisce il dizionario con nuovi esempi, come in questa sede accade per i nomi delle piante o degli animali, che, per questioni di spazio e obiettivi nel dizionario a stampa, non erano particolarmente approfonditi. Si è ritenuto che in un dizionario digitale si potesse dare più spazio agli esempi per chiarie l'uso e il significato delle parole legate alla natura e proprio tramite a questo attributo è possibile arricchire il dizionario attribuendo le responsabilità degli interventi. Come però è stato mostrato nel lemma campione al punto evidenziato in giallo, lo stesso ragionamento può essere ampliato per analogia ad infinite altre casistiche e interessi di studio. Il dizionario diventa così un luogo di confronto e un punto di incontro e va oltre alla sua funzione originaria di consultazione.

4. RESTITUZIONE DIGITALE

Il punto di arrivo del progetto è l'apertura della metodologia usata al contributo della comunità. Per la visualizzazione si può procedere alla trasformazione dall'XML al formato Markdown dal quale creare .pdf, .txt ed .epub, in modo che ad ogni elemento di contenuto corrisponda, quando necessario, un tipo di visualizzazione. In primo luogo, è prevista un'edizione a stampa di questo dizionario del latino ambientale, generata direttamente dal sorgente in XML TEI Lex-0. In secondo luogo, si prevede che il dizionario sia disponibile sul sito DigilibLT in modo interoperabile e integrato con la biblioteca digitale. Questo costituirà un prototipo che mostra come si possa colmare l'assenza di un dizionario latino-italiano recente, *online,* ad accesso aperto; e collaborativo perché si prevede la creazione di un ambiente digitale di lavoro dove, in modo controllato, si possa operare in modo supervisionato sui lemmi con un'attribuzione specifica della responsabilità.

Il risultato del progetto però consiste anche nella messa a punto di un processo di lavoro metodico replicabile per la digitalizzazione di altri dizionari, basato sull'analisi critica dei singoli passaggi, concepito per rispondere all'esigenza di produrre dati FAIR e caratterizzato da: utilizzo di risorse per quanto possibile aperte, collaborazione anche interdisciplinare, uniformazione a uno *standard* condiviso senza perdere la specificità del proprio testo.

BIBLIOGRAFIA

[1] Costa, Rute, Ana Salgado, Anas Fahad Khan, Sara Carvalho, Laurent Romary, Bruno Almeida, Margarida Ramos, Mohamed Khemakhem, Raquel Silva, e Toma Tasovac. «MOR Digital: The Advent of a New Lexicographical Portuguese Project». In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, (a cura di) Iztok Kosem, Michal Cukr, Miloš Jakubíček, Jelena Kallas, Simon Krek, e Carole Tiberius, 312–24. Lexical Computing CZ s.r.o., Brno, Czech Republic, 2021.

[2] Hallig, Rudolf, e Walther von Wartburg. *Begriffssystem als Grundlage für die Lexikographie / Système raisonné des concepts pour servir de base à la lexicographie*. Berlin: Akademie-Verlag, 1963.

[3] Mambrini, Francesco, Flavio Massimo Cecchini, Greta Franzini, Eleonora Litta, Marco Carlo Passarotti, e Paolo Ruffolo. «LiLa: Linking Latin. Risorse linguistiche per il latino nel Semantic Web». *Umanistica Digitale* 4, fasc. 8 (2020): 63–78. https://doi.org/10.6092/issn.2532-8816/9975.

[4] Romary, Laurent, e Toma Tasovac. «TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources». Tokyo, Giappone, 2018. https://inria.hal.science/hal-02265312.

[11] https://thesaurus.badw.de/tll-digital/tll-open-access.html

[5] Tiberius, Carole, Simon Krek, Katrien Depuydt, Polona Gantar, Jelena Kallas, Iztok Kosem, e Michael Rundell. «Towards the Elexis Data Model: Defining a Common Vocabulary for Lexicographic Resources». In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, (a cura di) Iztok Kosem, Michal Cukr, Miloš Jakubíček, Jelena Kallas, Simon Krek, e Carole Tiberius, 56–77. Lexical Computing CZ s.r.o., Brno, Czech Republic, 2021.

[6] Tittel, Sabine, Frances Gillis-Webber, e Alessandro A. Nannini. «Towards an Ontology Based on Hallig-Wartburg's Begriffssystem for Historical Linguistic Linked Data». In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, 1–10. European Language Resources Association, 2020. https://aclanthology.org/2020.ldl- 1.1.

LexiCad: piattaforma lessicografica digitale per l'italiano delle origini

Salvatore Arcidiacono[1], Antonella Zammataro[2]
[1] Università di Catania, Italia - salvatore.arcidiacono@unict.it
[2] Università di Catania, Italia - antonella.zammataro@phd.unict.it

ABSTRACT[1]

Il presente contributo descrive lo sviluppo di una piattaforma digitale per la lessicografia, denominata LexiCad.
Dopo un inquadramento generale su alcuni problemi connessi alla lessicografia elettronica storica in Italia, verrà illustrata sinteticamente l'architettura del sistema LexiCad, che si propone come un sistema flessibile e interoperabile per la creazione di piattaforme lessicografiche *web-based*. Tra i risultati del lavoro, verranno forniti i riferimenti ai progetti in cui questa tecnologia è stata impiegata; verranno inoltre delineati i vantaggi strategici che l'adozione di un modello simile potrebbe apportare in termini di interoperabilità e ottimizzazione delle risorse.

PAROLE CHIAVE

LexiCad; lessicografia; DWS.

1. INTRODUZIONE: INFORMATICA E LESSICOGRAFIA STORICA

I sistemi informativi impiegati in lessicografia si distinguono, tra quelli in uso nelle discipline umanistiche, per le dimensioni molto ampie: le ricerche di lessicografia diacronica – che nella classificazione dell'infrastruttura europea ELEXIS[2] possono essere collocate nel contesto della *Professional Large-scale Lexicography* [9: 887-888] – prendono infatti in analisi un'ingente quantità di dati testuali (le fonti primarie oggetto di spoglio) e producono risultati altrettanto complessi, caratterizzati da una macrostruttura estesa e da una microstruttura particolarmente ricca. L'informatizzazione ha reso notevolmente più efficienti le procedure ma, anche a causa della crescente affermazione delle metodologie *corpus-based*, la gestione dell'informazione continua a costituire un fattore determinante per il successo delle ricerche e a condizionare l'esaustività delle attività di spoglio e schedatura.
In Italia, il progetto per un *Vocabolario Storico*, avviato per iniziativa dell'Accademia della Crusca e affidato oggi al Consiglio Nazionale delle Ricerche, dal 1972 ha concentrato i suoi sforzi sul *Tesoro della Lingua Italiana delle Origini* (*TLIO*), costruito sullo spoglio integrale di un *corpus* che mira a raccogliere l'intera produzione volgare scritta fino al limite cronologico del 1375, con alcuni casi di sconfinamento agli inizi del XV secolo[3]. Accanto al *Corpus TLIO*, lemmatizzato e costituito da quasi ventiquattro milioni di occorrenze, l'OVI ha allestito anche il *Corpus OVI dell'italiano antico*, non lemmatizzato ma più ampio. Per gestire queste risorse, l'OVI ha sviluppato il programma GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini), progettato e realizzato negli anni Novanta dal compianto Domenico Iorio-Fili. Le diverse versioni del *software*[4] sono state adottate da un numero crescente di progetti, raccogliendo attorno a GATTO un ricco insieme di iniziative scientifiche che, condividendo lo strumento, convergono sulle procedure e sui modelli definiti dall'OVI e, di conseguenza, pervengono a risultati scientifici confrontabili.
Mentre con GATTO la lessicografia storica italiana ha potuto disporre di un eccellente *Corpus Query System* (CQS), si è a lungo sentita la necessità di un sistema altrettanto avanzato di redazione e pubblicazione (*Dictionary Writing System*, o DWS)[5]. Le ricerche volte a sviluppare un DWS per l'italiano delle origini, per la verità, vennero avviate già negli anni Novanta del secolo scorso, periodo in cui si affermò l'uso di questo tipo di *software* presso i principali cantieri lessicografici, con il progetto COVIRED[6]. La procedura COVIRED, purtroppo, non riuscì a entrare in funzione e per

[1] Il presente contributo è il risultato di un costante confronto fra gli autori nelle fasi di ideazione, stesura e revisione: tuttavia, la responsabilità dell'introduzione va attribuita ad A. Zammataro, quella dei paragrafi 2 e 3 a S. Arcidiacono.
[2] https://elex.is
[3] Come prescrivono le Norme di Beltrami [3: 113], il *Corpus TLIO* rappresenta il nucleo di informazione primario per il redattore, il quale però, con parsimonia, può ampliare la propria ricerca ad altre fonti ritenute pertinenti. Il *corpus*, in questo contesto, più che un limite agli spogli, è da intendersi come una garanzia di completezza [2: 240].
[4] In particolare GATTO 3.3, rilasciato pubblicamente dall'Istituto, e la sua controparte destinata alla consultazione in rete GattoWeb.
[5] Il DWS costituisce il sistema che assiste la scrittura delle voci e che ne gestisce la rappresentazione digitale, la pubblicazione in rete e l'archiviazione.
[6] La procedura COVIRED, di cui rimangono solo alcune descrizioni sommarie, comprendeva anche lo sviluppo di un CQS per il *corpus* che si andava costituendo per il *TLIO* [7]. Il DWS previsto dal sistema mirava a definire una procedura di redazione fortemente informatizzata ma finalizzata alla pubblicazione a stampa e su CD-ROM.

l'avvio del *TLIO* si è scelto di utilizzare un comune *word processor*[7]. Nel 2011, presso l'OVI-CNR, è stato avviato lo studio e lo sviluppo di un vero e proprio DWS volto a sostituire la redazione delle voci in Word, denominato ReddiX[8], descritto in pochi articoli [4] [5] e in alcune preziosissime relazioni tecniche interne. Purtroppo, neppure questo progetto ebbe fortuna: intorno al 2015, quando il progetto ReddiX fu abbandonato, le risorse digitali raccolte dal CNR-OVI avevano raggiunto dimensioni ragguardevoli e una complessità tale da richiedere un notevole livello di personalizzazione degli strumenti, scoraggiando quindi il ricorso a qualsiasi *software* di terze parti. Tuttavia, nell'ambito del progetto ReddiX fu realizzato un convertitore capace di trasformare in XML le voci redatte Word in XML[9] che, oltre a scongiurare i seri rischi di obsolescenza del formato .doc, ha rappresentato il punto di partenza per la definizione di un nuovo sistema.

2. LA PIATTAFORMA LEXICAD

Nel corso dei lavori preparatori per il *Vocabolario del Siciliano Medievale* (*VSM*, diretto da Mario Pagano presso l'Università di Catania in concorso con il Centro di studi filologici e linguistici siciliani), è stato sviluppato il sistema LexiCad, una sorta di '*framework'* lessicografico per lo sviluppo di dizionari elettronici: come nel caso dei *framework* utilizzati nello sviluppo *software*, LexiCad si configura come uno strumento per accelerare la realizzazione di nuove piattaforme lessicografiche digitali. In tal senso, questo sistema, costruito sullo *stack* LAMP[10], implementa un *design pattern*, stabilisce un insieme di convenzioni uniformi ed estende il linguaggio PHP, riproponendo in un oggetto *software* i modelli in uso nella lessicografia dell'italiano delle origini, secondo i criteri elaborati presso l'Opera del Vocabolario Italiano [1].

L'architettura di un progetto basato su LexiCad è organizzata su diversi livelli di astrazione. A quello più basso sono state raccolte le funzioni essenziali per il funzionamento di qualsiasi piattaforma *web*[11]. A un livello intermedio è stato ricondotto tutto ciò che pertiene al trattamento di dati linguistici e filologici. Nel concreto, questo strato si compone di classi astratte che descrivono e rendono computabili le entità di base più comuni nel trattamento di informazioni testuali[12]. Una classe, per esempio, è dedicata alle forme grafiche, sia che si tratti delle forme di una voce[13], sia delle forme di un atlante linguistico o di quelle di un *corpus*: la classe fornisce allo sviluppatore gli strumenti per codificare e archiviare le forme, per allestire e ordinare formari di varia natura, per gestire l'omografia, per trovare i lemmi a cui una forma è associata, ecc. Molto più complessa è invece la classe dedicata alle voci: questa ha la funzione di gestire la microstruttura nei suoi componenti più elementari ma anche di organizzare dinamicamente la macrostruttura del vocabolario[14]; questa classe si occupa, tra l'altro, di istanziare e coordinare gli oggetti relativi ad altre entità coinvolte nella microstruttura (categorie grammaticali, forme, etimi, accezioni, esempi, marche d'uso, marche grammaticali, ecc.). L'ultimo strato, il più alto nella gerarchia, è dedicato alle personalizzazioni della singola piattaforma, e consente di aggiungere funzioni a quelle già esistenti o di estendere le classi degli strati inferiori. Con l'eccezione del livello più profondo, LexiCad è organizzato in 'applicazioni' che possono essere spostate da un'implementazione all'altra.

Dopo la prima versione della piattaforma per il *VSM* e un adattamento per l'*Atlante Grammaticale dell'Italiano delle Origini* (*AGLIO*)[15], questa tecnologia è stata adottata dal CNR-OVI attraverso lo sviluppo di Pluto (Piattaforma

[7] Nel 1996 la redazione delle voci del *TLIO* è stata avviata ricorrendo al programma di videoscrittura Microsoft Word. I *file* .doc con le voci erano rigorosamente strutturati per mezzo di un foglio di stile che ha permesso, già dall'anno successivo, lo sviluppo di un *parser* per la conversione in HTML. Le voci in HTML erano poi trasferite su un sistema di pubblicazione in rete, basato su un unico *file* (appena sufficiente a contenere le prime voci) indicizzato da uno *script* in PERL in ambiente SunOS [5: 28]. Una simile architettura, come previsto, non garantiva una sufficiente scalabilità e, nel 2002, fu sostituita dal TLIOWeb, il sistema che fino a oggi ha permesso la consultazione del *TLIO* in rete. Il TLIOWeb non è un DWS, ma una «soluzione 'ponte'» [4: 57], cioè un complesso insieme di procedure e di applicazioni che permettono di trasformare e pubblicare sul web le voci del vocabolario. Terminata la conversione delle oltre cinquantamila voci redatte in Word, grazie a un convertitore implementato con LexiCad, Pluto prenderà il posto del sistema TLIOWeb anche come strumento di consultazione del *TLIO* in rete.

[8] La 'X' finale nel nome indica la centralità del *markup* XML nell'architettura del sistema. ReddiX era composto da due 'macromoduli' – costituiti da due programmi separati che interagivano secondo una logica *client-server* – uno *off-line* per la redazione delle voci e uno *on-line* per la pubblicazione in rete e a stampa del dizionario.

[9] I *file* XML erano validati su una DTD (*Document Type Definition*) modellata sulle specifiche caratteristiche dei dati estratti.

[10] L'acronimo si riferisce, come noto, all'ambiente Linux, al server web Apache, al linguaggio PHP e al server per i *database* MySQL.

[11] Questo livello fornisce anche le funzioni relative alla gestione delle pagine, degli utenti, dei permessi e dei *backup* periodici, ecc.

[12] Il termine 'classe' è qui inteso nei termini del costrutto in uso nella programmazione orientata agli oggetti.

[13] I dizionari che si fondano sul modello del *TLIO* prevedono, nel punto 0.1 dell'intestazione della voce, un elenco delle forme grafiche attestate nella documentazione.

[14] Questa classe può generare diversi tipi di indici ricercabili, tenendo conto, per esempio, dei rapporti di derivazione e composizione tra i lessemi, dei rinvii o di una qualsiasi condizione basata su uno dei campi previsti dalla voce.

[15] Diretto da Marcello Barbato e Vincenzo Faraoni; http://aglio.ovi.cnr.it

Lessicografica Unica del Tesoro delle Origini)[16]. Gli obiettivi di questa operazione non si limitano al trasferimento su una nuova piattaforma delle oltre quarantamila voci del *TLIO*, ma prevedono l'integrazione in un unico sistema dei numerosi *asset* digitali dell'Istituto, con l'eccezione dei *corpora* in GATTO[17].
Dal punto di vista dell'utente, il sistema LexiCad / Pluto è composto da un *back-end* di redazione e di amministrazione, riservato ai redattori e agli utenti autorizzati, e da un *front-end* di consultazione ad accesso pubblico. Il cuore del *back-end* è la maschera di redazione della voce (vd. Fig. 1), in cui la microstruttura del vocabolario viene mappata su specifici componenti dell'interfaccia (*form*, *editor* testuali WYSIWIG, menu di scelta multipla, ecc.).

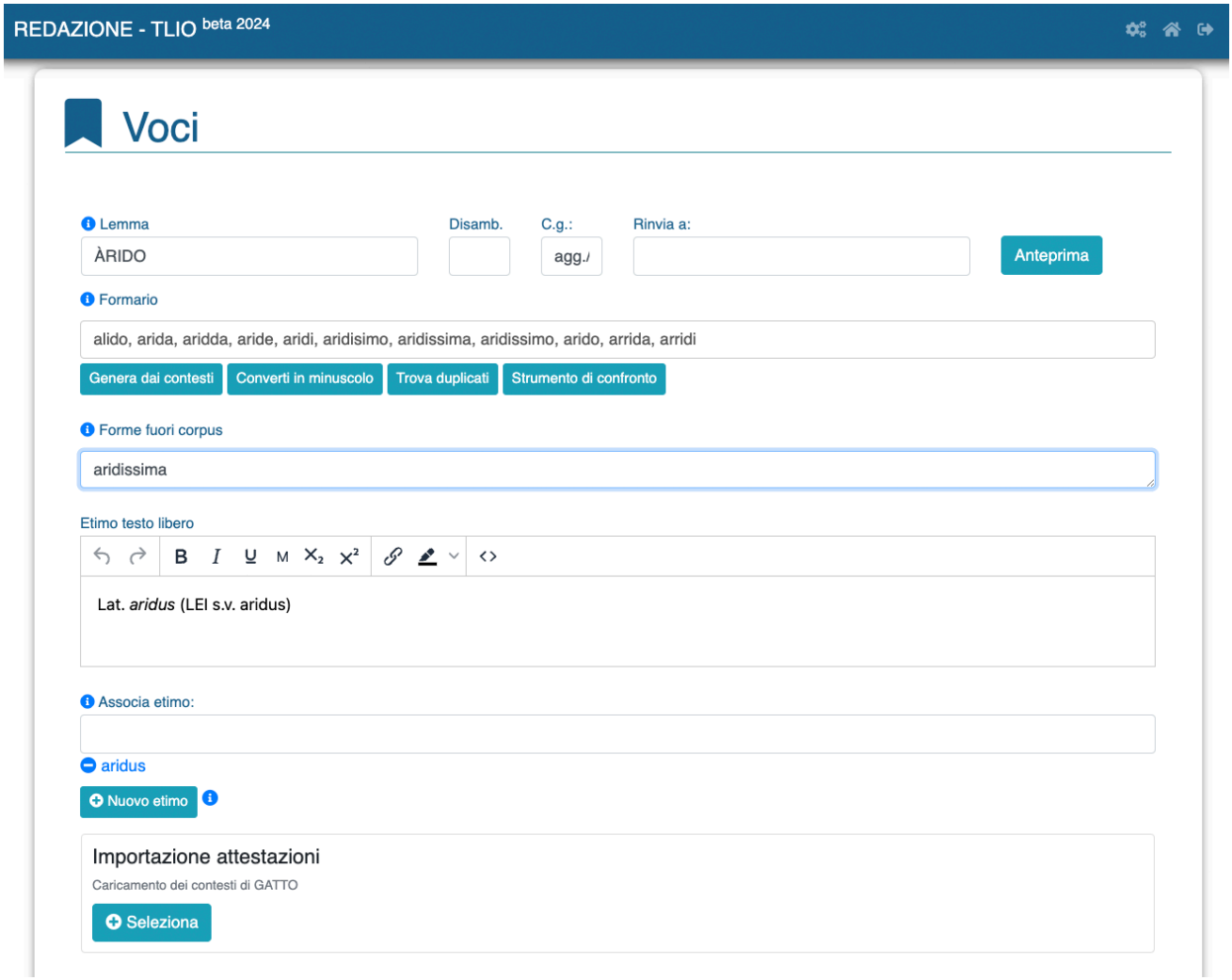


Figura 1. La maschera di redazione

In qualsiasi punto della maschera è possibile incapsulare *script* dedicati o vere e proprie applicazioni personalizzate. Alcuni punti della voce possono essere omessi e sostituiti da procedure di compilazione automatica; in altri casi possono essere approntati in modo semi-automatico grazie ad algoritmi in grado di accedere a tutti gli archivi di Pluto o a servizi esterni. I riferimenti mesostrutturali – che includono i collegamenti tra entrate del vocabolario o i loro costituenti così come qualunque riferimento a entità presenti sulla piattaforma – sono istituiti a partire da caselle di ricerca e di selezione per garantire una maggiore coerenza redazionale e l'integrità referenziale del complesso informativo. Tutti gli elementi a

[16] La gran parte dell'attività di sviluppo su Pluto è oggi condotta nell'ambito del progetto *QM (Quattrocento Meridionale) - Il futuro dell'italiano antico* (PRIN 2020).
[17] Per riprendere l'esempio dell'ATILF (Analyse et Traitement Informatique de la Langue Française), l'implementazione di un sistema all'interno di un contesto istituzionale come quello del CNR-OVI può contribuire a promuovere quella complementarità 'interna' tra progetti scientifici, così come definita da Buchi [6: 5], che si ottiene con l'istituzione di relazioni sistematiche tra le ricerche, fino a costruire un unico «dispositivo lessicografico (e metalessicografico) ragionato, dove ogni risorsa ha il suo ruolo». Valutazioni simili hanno guidato lo sviluppo della piattaforma digitale Pasadena dell'*OED* (*Oxford English Dictionary*), così come documentato Elliot e Williams [8: 258].

inserimento guidato o vincolato (caselle di scelta multipla, suggerimenti automatici, ricerca guidata, ecc.) concorrono ad accelerare ulteriormente la redazione della voce.
Il pannello di redazione è concepito come un ambiente collaborativo: ciascuna implementazione di LexiCad stabilisce le proprie politiche di accesso alle singole risorse, definendo rigorosamente le operazioni che un utente è autorizzato a compiere sui contenuti (lettura, inserimento, scrittura o cancellazione). Il sistema dei permessi è strettamente connesso agli strumenti di gestione del flusso di lavoro: per esempio, lungo le fasi previste di redazione e revisione di una voce, il sistema consente o inibisce l'accesso alle voci o ad alcuni punti della microstruttura in relazione alle figure chiamate in causa in uno specifico momento del flusso redazionale[18].

3. SVILUPPI E PROSPETTIVE

L'architettura della piattaforma ha reso estremamente facile il riadattamento di LexiCad a nuove implementazioni che si possono collocare in differenti contesti di ricerca applicata. Questa tecnologia, oltre a essere utilizzata per il trattamento e la riconversione delle voci del *TLIO*, gestisce attualmente altri quattro progetti lessicografici specializzati: il già citato *VSM*; il *Vocabolario Dantesco* (*VD*, diretto da Paola Manni e Lino Leonardi)[19], il *Vocabolario Dantesco Latino* (*VDL*, diretto da Gabriella Albanese, Paolo Chiesa e Mirko Tavoni)[20] e il *Vocabolario storico-etimologico del Veneziano* (*VEV*, diretto da Lorenzo Tomasin e Luca D'Onghia)[21]; sono in corso di sviluppo, inoltre, le piattaforme per il *Dizionario Etimologico e Storico del Napoletano* (*DESN*, diretto da Nicola De Blasi e Francesco Montuori) e per il *Vocabolario del romanesco contemporaneo* (*VRC*, diretto da Paolo D'Achille e da Claudio Giovanardi)[22].
Il sistema di redazione è indipendente dal CQS ma può essere agevolmente integrato con diversi *software* di interrogazione attraverso un gestore dei contesti: Pluto dispone di un'interfaccia in grado di importare da GATTO 4 un *file* XML con i risultati delle ricerche corredati da una serie di informazioni supplementari; la piattaforma per il *VSM* ha introdotto il supporto ai *file* in 'formato redazionale' che possono essere esportati da GattoWeb; la piattaforma del *VDL* è stata collegata a *Dante Search*[23] attraverso un'API (*Application Programming Interface*) sviluppata dall'ISTI-CNR. I diversi progetti possono quindi mantenere il motore di interrogazione del proprio *corpus* ma i dati potranno essere comunque agevolmente trasferiti all'interno dell'interfaccia. Effettuato il trasferimento, la maschera di redazione consentirà al redattore di interrogare, modificare e annotare i contesti da riportare sotto le accezioni del dizionario.
Oltre alle implementazioni propriamente lessicografiche, LexiCad è stato impiegato in altri progetti che, a vario titolo, si collocano nell'orizzonte delle *digital humanities*, come nel caso dei progetti *ItalArt* (*L'italiano delle arti tra Medioevo e Rinascimento*)[24], *RdP* (*Rime disperse di Petrarca*)[25], il sito del Gruppo Guiron per il *Ciclo di Guiron le Courtois*[26] e la *Bibliografia dei Commenti Danteschi* (*BCD*) del CNR-OVI[27]. Alla georeferenziazione dei dati linguistici è dedicato il modulo LexiMap, approntato per gestire le carte dell'*Atlante dell'Italiano delle Origini* (*AGLIO*) su impulso del progetto MIRA (Mappatura dell'Italo-Romanzo Antico), dedicato alla mappatura geolinguistica dei volgari italoromanzi[28]. In seguito, le tecniche di proiezione dei dati geolinguistici su mappe *web* sono state riprese nell'ambito di uno studio in collaborazione con il Notre Dame Center for Italian Studies della University of Notre Dame, con il fine di generalizzare l'applicazione di questi metodi, estendendoli dall'*AGLIO* a tutti gli studi sull'italiano delle origini. Nell'immediato futuro, Pluto potrebbe implementare i risultati di queste ricerche sulla *Bibliografia dei Testi Volgari* (già predisposta ad accogliere le coordinate delle singole aree) e, in un secondo momento, utilizzare i metadati delle schede bibliografiche per geolocalizzare alcuni punti delle voci del *TLIO*[29].

[18] Il sistema dei permessi consente anche di inserire nella microstruttura della voce una serie punti 'di servizio' non accessibili all'utente finale, come nei casi delle annotazioni dedicate ai processi di revisione.
[19] http://www.vocabolariodantesco.it
[20] http://www.vocabolariodantescolatino.it
[21] http://vev.ovi.cnr.it
[22] LexiCad è un *software* proprietario ma le sue implementazioni sono realizzate nell'ambito di accordi stipulati con l'OVI-CNR o con l'Università di Catania.
[23] https://dantesearch.dantenetwork.it
[24] http://italart.ovi.cnr.it
[25] http://rdp.ovi.cnr.it
[26] https://guiron.fefonlus.it
[27] http://bcd.ovi.cnr.it
[28] Il progetto MIRA è diretto da Michele Loporcaro presso l'Università di Zurigo e finanziato dal Fondo Nazionale Svizzero; https://data.snf.ch/grants/grant/205028.
[29] http://pluto.ovi.cnr.it/btv

I benefici relativi alla condivisione di un unico sistema riguardano principalmente l'interoperabilità[30] delle piattaforme, la standardizzazione dei processi redazionali e la confrontabilità dei risultati dei diversi progetti. È possibile osservare che la scelta di costruire un *framework* per la realizzazione di dizionari digitali, invece che un singolo *software*, ha prodotto un abbassamento degli sbarramenti tecnici che costituisce, di per sé, un'ulteriore misura dell'efficacia di questo approccio: i contesti in cui la piattaforma è stata riadattata sono iniziative medio-piccole che, negli studi di fattibilità che precedono l'avvio di un nuovo progetto lessicografico (specialmente nei progetti competitivi), sarebbero state penalizzate dalla carenza di risorse necessarie per lo sviluppo di un *software* dedicato.
Questo approccio asseconda inoltre quell'orientamento alla specializzazione lessicografica che si riscontra nella lessicografia storica moderna: il *framework* metodologico e tecnologico permette di intraprendere agili analisi lessicografiche, concentrate su specifiche aree geolinguistiche o singoli autori, pur mantenendo aperta la possibilità di integrare le risorse lessicali attraverso piattaforme interoperabili.

BIBLIOGRAFIA

[1] Arcidiacono, Salvatore. *Lessicografia elettronica e italiano delle origini*. Palermo: Centro di studi filologici e linguistici siciliani, 2022.

[2] Beltrami, Pietro. «Lessicografia e filologia in un dizionario storico dell'italiano antico». In *Storia della lingua e filologia. Atti del convegno ASLI (Pisa-Firenze, 18-20 dicembre 2008)*, (a cura di) Claudio Ciociola, 235–248. Cesati, 2010.

[3] Beltrami, Pietro ((a cura di)). *Norme per la redazione del Tesoro della Lingua Italiana delle Origini*, 1998. http://tlio.ovi.cnr.it/TLIO/NormeTLIO.pdf.

[4] Boccellari, Andrea. «Il sistema di redazione e pubblicazione web del TLIO». In *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori (Firenze, 26 ottobre 2010)*, 57–64. Supplementi al Bollettino dell'Opera del Vocabolario Italiano, 3. Edizioni dell'Orso, 2012.

[5] Boccellari, Andrea, e Domenico Iorio-Fili. «Il supporto dell'informatica al Vocabolario». In *L'Opera del Vocabolario Italiano per Pietro G. Beltrami*, (a cura di) Paolo Squillacioti e Giulio Vaccaro, 15–30. Edizioni dell'Orso, 2013.

[6] Buchi, Eva. «La lessicografia storica condotta dall'ATILF: ancoraggio lessicologico, complementarietà interna e internazionalità crescente». In *L'italiano dei vocabolari. Atti della sesta edizione della Piazza delle Lingue*, (a cura di) Nicoletta Maraschio, Domenico De Martino, e Giulia Stanchina, 3–10. La Piazza delle lingue, 4. Accademia della Crusca, 2013.

[7] Ceccoli, Ubaldo, Franco Lorenzi, e Valentina Pollidori. «Un programma per la redazione del Vocabolario Storico della Lingua Italiana assistita dal calcolatore». In *Récit et Informatique. Actes de la journée d'études*, (a cura di) Claude Cazalé Berard, 67–84. Editions de l'Espace Européen, 1989.

[8] Elliott, Laura, e Sarah Williams. «Pasadena. A New Editing System for the Oxford English Dictionary». In *Atti del XII Congresso Internazionale di Lessicografia - Proceedings of the XII EURALEX International Congress*, (a cura di) Elisa Corino, Carla Marello, e Cristina Onesti, 257–264. Edizioni dell'Orso, 2006.

[9] Krek, Simon, Kosem Iztok, John P. McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, e Tanja Wissik. «European Lexicographic Infrastructure (ELEXIS)». In *Global Contexts. Proceedings of the XVIII EURALEX International Congress*, 881–891. Ljubljana University Press, 2018.

[30] I principali moduli del sistema sono corredati da un *set* di API che permettono di scambiare dati in formato JSON e XML. Inoltre, sarà presto rilasciato un sistema di esportazione delle voci codificate in XML/TEI, per il quale si rimanda al contributo di Giuseppe Zappalà in questo stesso volume.

Strategie per la creazione e la condivisione di una collezione digitale di testi greco-latini

Vincenzo Ortoleva[1], Maria Rosaria Petringa[2], Salvatore Cammisuli[3], Mariarosaria Villareale[4]
[1]Università di Catania, Italia - ortoleva@unict.it
[2]Università di Catania, Italia - mrpetri@unict.it
[3]Università di Catania, Italia - salvatore.cammisuli@unict.it
[4]Università di Catania, Italia - mariarosaria.villareale@virgilio.it

ABSTRACT

Il contributo presenta le strategie volte alla creazione e alla condivisione di una collezione digitale di testi greco-latini, intitolata *Onomastikón, Studi di lessicografia greca e latina* (https://onomastikon.altervista.org/), costantemente ampliata e aggiornata attraverso l'interazione tra docenti/ricercatori strutturati e studenti del Corso di Laurea in "Filologia classica" dell'Università di Catania, nonché attraverso la collaborazione con istituzioni scientifiche europee ed extra-europee, quali il *Thesaurus linguae Latinae* della Bayerische Akademie der Wissenschaften di Monaco di Baviera (Germania), il *Diccionario Griego-Español* dell'Instituto de Lenguas y Culturas del Mediterráneo y Oriente Próximo - Centro de Ciencias Humanas y Sociales - CSIC di Madrid (Spagna) e con il *Thesaurus Linguae Graecae*® (*TLG*)® della University of California, Irvine (Stati Uniti d'America). In particolare, tale collezione ha come frutto da una parte la pubblicazione di porzioni, precedentemente inedite, di un glossario bilingue latino-greco, risalente alla tarda antichità. Tali testi sono forniti in edizione critica commentata, su una piattaforma *online* e in *open access*. A ciò si affianca dall'altra il progetto di un *Archivio digitale del lessico della poesia cristiana antica e medievale*, che mira all'individuazione e archiviazione in *open access* di una cospicua serie di termini poco o per nulla altrove attestati, che si rinvengono nei testi poetici latini cristiani di età tardoantica, al fine di creare una banca dati aggiornata del lessico specialistico della poesia cristiana antica finora mai attuata. È inoltre in corso l'implementazione di raffinate risorse di ricerca, che permetteranno agli studiosi interessati di compiere molteplici indagini lessicali.

PAROLE CHIAVE

Glossari; edizione critica in *open access*; interazione tra ricercatori strutturati e studenti; poesia cristiana latina; archivio digitale.

1. INTRODUZIONE

Il progetto *Onomastikón, Studi di lessicografia greca e latina*, coordinato da Vincenzo Ortoleva, Maria Rosaria Petringa e Salvatore Cammisuli, ha nel suo complesso l'obiettivo di creare interazione tra docenti/ricercatori strutturati e studenti universitari nel campo delle scienze dell'antichità. Tale rete si propone di contribuire al ripensamento delle attività e dei metodi relativi alla didattica e alla ricerca, sfruttando le opportunità offerte dagli strumenti digitali. Il progetto si interfaccia altresì con istituzioni scientifiche europee ed extra-europee, quali il *Thesaurus linguae Latinae* della Bayerische Akademie der Wissenschaften di Monaco di Baviera (Germania), il *Diccionario Griego-Español* dell'Instituto de Lenguas y Culturas del Mediterráneo y Oriente Próximo - Centro de Ciencias Humanas y Sociales - CSIC di Madrid (Spagna) e con il *Thesaurus Linguae Graecae*® (*TLG*)® della University of California, Irvine (Stati Uniti d'America).

La rete di lavoro ha un duplice campo di indagine. Il primo riguarda la lessicografia greca e latina, con l'obiettivo specifico di produrre edizioni critiche commentate di porzioni del glossario latino-greco dei cosiddetti *Hermeneumata Celtis*. I risultati di tali ricerche sono condivisi in formato digitale e *open access* a partire dal febbraio 2020 sul sito https://onomastikon.altervista.org. Le sezioni, o le porzioni di sezioni, ivi pubblicate sono il risultato del lavoro delle tesi di Laurea magistrale di alcuni studenti del Corso di Laurea in "Filologia classica" dell'Università di Catania.

A ciò si affianca, a pari titolo, il progetto di un *Archivio digitale del lessico della poesia cristiana antica e medievale*. Esso è in fase di implementazione e mira ad allestire mediante vari *step*, a partire dagli studi preliminari di carattere linguistico e critico-testuale e dal lavoro di raccolta dati fino alla sua messa in opera definitiva tramite *open access*, passando attraverso le fasi intermedie di programmazione e prova, un *database* aggiornato del lessico specialistico della poesia cristiana antica – finora mai attuato – dei termini poco o per nulla altrove attestati, che si rinvengono nei testi poetici latini cristiani tardoantichi. Di tali particolarità lessicali si chiarisce inoltre il significato e si delineano anche gli sviluppi nel latino medievale oltre che i possibili esiti romanzi. A tal fine sono utili anche i confronti con i lessici e i glossari tardoantichi, sia quelli editi che inediti, compreso quello degli *Hermeneumata Celtis*.

2. STATO DELL'ARTE

Gli *Hermeneumata Celtis*, che costituiscono una delle redazioni dei cosiddetti *Hermeneumata Pseudodositheana*, sono tramandati dal cod. Wien, Österreichische Nationalbibliothek suppl. Gr. 43. Tale manoscritto, cartaceo, comprende due parti originariamente separate. Nella prima (ff. 1v-11v) si rinviene una grammatica greca, rielaborazione di materiale ricavato da analoghi testi di epoca umanistica: essa è trascritta da Johannes Rosenberger, copista personale dell'umanista tedesco Conrad Celtis (1459-1508). La seconda contiene appunto gli *Hermeneumata*, copiati dalla mano dello stesso Celtis, dal quale prendono il nome.

Conformemente all'impostazione di questo tipo di manuali, anche gli *Hermeneumata Celtis* si articolano a loro volta in due distinte sezioni: un *colloquium* bilingue greco-latino, costituito da semplici dialoghi, finalizzati all'apprendimento delle due lingue (ff. 12-17), e un glossario latino-greco, organizzato in sezioni tematiche (ff. 18-45v). La parte relativa agli *Hermeneumata* fu copiata da Celtis nel 1495 da un manoscritto, oggi purtroppo perduto, da lui ritrovato nell'abbazia benedettina di Sponheim, nella Renania-Palatinato (Germania).

Al momento dell'implementazione del progetto, gran parte del glossario risultava ancora del tutto inedita. Delle cinquanta sezioni tematiche in cui è articolato il glossario, solamente otto erano state pubblicate, in varie sedi, da autori diversi [10-13]. A tutt'oggi il portale *Onomastikón* ha messo a disposizione ben ventuno tesi di laurea, per un totale di ventisei sezioni completate.

Per quanto riguarda l'archivio digitale del lessico specialistico della poesia cristiana antica e medievale, una cospicua serie di dati provenienti dalla schedatura delle opere in versi è stata approntata e confluirà in tale banca dati attraverso progressivi aggiornamenti, accompagnata da un'ampia bibliografia scientifica di riferimento. Allo stato attuale sono stati prodotti innovativi studi relativi all'indagine delle particolarità linguistiche rinvenibili nelle opere poetiche cristiane latine considerate nel loro precipuo genere letterario. In particolare, con riferimento, ad esempio, ai fenomeni di risemantizzazione, riuso del lessico e degli stilemi della poesia classica e creazione di neologismi, sono stati già pubblicati numerosi contributi dalla Responsabile scientifica del progetto dell'*Archivio digitale del lessico della poesia cristiana antica e medievale*, cf. [17-24, 25, 29, 30-40].

3. STRATEGIE ADOTTATE

Nel contesto attuale continuano a perdurare alcune abitudini acquisite durante la recente emergenza pandemica e ancora non del tutto superate: di fatto, in assenza di azioni mirate, il confronto e il dialogo tra studenti è spesso alquanto ridimensionato, con il rischio che ciascuno di essi si chiuda nel proprio lavoro personale, limitandosi all'interazione con il docente. Senza l'intervento di un'adeguata formazione mirata, inoltre, la stessa Rete Internet, che è ormai la principale fonte di informazione dello studente, non è esente da rischi quanto all'affidabilità e alla scientificità dei contenuti in essa presenti.

Nella creazione di contenuti scientificamente qualificati il progetto coinvolge attivamente gli studenti. Questo è un aspetto assolutamente innovativo e raramente presente nel panorama della ricerca italiana. La portata innovativa del progetto emerge dunque soprattutto dal contributo al ripensamento delle metodologie relative alla didattica e alla ricerca. Gli studenti, secondo l'approccio educativo del *peer-tutoring*, si scambiano conoscenze, competenze ed esperienze, contribuendo attivamente alla costruzione delle metodologie e delle strategie operative del gruppo di lavoro. Tale interazione avviene principalmente nel corso di incontri, tenuti *online* e/o in presenza, a ciascuno dei quali partecipa un ricercatore strutturato in qualità di facilitatore.

Relativamente al progetto di una prima edizione critica del glossario degli *Hermeneumata Celtis*, sono stati coinvolti in primo luogo gli studenti del Corso di laurea magistrale in "Filologia classica" del Dipartimento di Scienze Umanistiche dell'Università degli Studi di Catania, ma la rete è comunque aperta alla partecipazione di studenti di altre Università italiane e internazionali. A queste ricerche è connaturato un approccio concretamente interdisciplinare: lo studio di un glossario latino-greco, i cui contenuti sono spesso legati alla cultura materiale, richiede infatti non solo la piena integrazione dello studio delle due lingue antiche, ma anche il vaglio più ampio e differenziato possibile delle varie fonti (siano esse letterarie, epigrafiche o archeologiche).

Per quanto riguarda l'aspetto operativo, ottenuta la liberatoria da parte dell'autrice o dell'autore della tesi, ciascun documento, non appena pronto, è di volta in volta pubblicato sul portale, dove i singoli lavori sono raggruppati per anno. Va messo in evidenza che ogni annata è dotata di un proprio codice ISBN. Si è posto altresì il problema della strategia da adottare al fine di ottenere, da un lato, la massima pubblicità dei contenuti, dall'altro, la loro tutela. Per raggiungere tale duplice fine, i documenti in formato pdf sono protetti da una password, liberamente concessa agli studiosi che ne fanno richiesta.

La consultazione del *corpus* è al momento consentita dall'indicizzazione del sito da parte di Google, con relativa funzione di ricerca interna, presente nel portale. Tale funzione, dopo un periodo di improvvisa interruzione, tornerà presto pienamente operativa. È inoltre in corso l'implementazione di più sofisticate risorse di ricerca, che permetteranno agli studiosi interessati di compiere indagini più raffinate: ricerche di contesti; ricerca per forme, per lemmi, per categorie grammaticali e per disambiguatori; cooccorrenze; generazione di *indices locorum*; produzione di liste di dati; esportazione di informazioni statistiche.

4. RISULTATI OTTENUTI E ATTESI

Come si è già accennato, al momento ben ventitré sezioni del glossario degli *Hermeneumata Celtis*, sulle complessive cinquanta, sono state pubblicate, in edizione critica commentata, sul portale *Onomastikón*. Nel medio-lungo termine si attende un ulteriore e significativo progresso: già allo stato attuale numerose nuove sezioni del glossario risultano assegnate e/o in lavorazione a studenti universitari, in vista del conseguimento del proprio titolo di Laurea magistrale.

Il progetto *Onomastikón* è stato presentato in prestigiose sedi internazionali, ottenendo l'interesse e l'apprezzamento di studiosi e istituzioni scientifiche italiane e straniere, che finalmente possono giovarsi di una collezione – non solo *online* ma anche in *open access* – di testi greco-latini precedentemente inediti. Prova della validità dei lavori prodotti è il fatto che tesi pubblicate su *Onomastikón* sono citate e discusse in studi apparsi o che stanno per essere pubblicati su riviste scientifiche a diffusione internazionale: cf. [1-3, 9, 15, 26-28].

Inoltre, per quanto riguarda l'*Archivio digitale del lessico della poesia cristiana antica e medievale*, varie tesi di laurea magistrale sull'argomento sono state già discusse, e altre risultano assegnate e in lavorazione, contribuendo a porre problematicamente in luce l'interpretazione corretta di specifici passi. Molti dei risultati frutto delle indagini scientifiche sono stati già presentati in vari seminari nazionali e convegni internazionali, i cui atti sono pubblicati o in corso di pubblicazione, cf. [17-24, 25, 29, 30-40]. Si prevede infine di organizzare un Convegno su tematiche specifiche del latino tardoantico al fine di rendere più proficuo il confronto sul piano scientifico e metodologico con studiosi delle Istituzioni scientifiche non solo di area mediterranea, tra cui anche quelle summenzionate già aderenti al Progetto.

Le attività di ricerca degli studenti risultano valorizzate e arricchite: oltre allo sviluppo di competenze informatiche, mediante la messa a punto di piattaforme digitali dedicate, nel contesto dell'innovazione delle *digital humanities*, gli studenti acquisiscono *soft skill* spendibili nel mondo del lavoro, quali *team building*, flessibilità, attitudine all'inclusività, all'apprendimento continuo e al pensiero critico.

Ancora, obiettivo precipuo del progetto è la sua ecosostenibilità, che si concretizza in un'ottica autenticamente *green*: un adempimento doveroso nei confronti delle nuove generazioni e del pianeta. La creazione di contenuti esclusivamente digitali permette, infatti, per via dell'assenza di materiale cartaceo, la minimizzazione dell'impatto ambientale e il contrasto al consumo di risorse.

Infine, per ottenere un'ancor più ampia condivisione della collezione completa di testi, è in programma il caricamento di tutto il materiale sul portale *Internet Archive*. Ciò garantirà, altresì, la conservazione nel tempo di tali contenuti digitali.

BIBLIOGRAFIA

[1] Cammisuli, Salvatore. «La sez. 11 del glossario degli *Hermeneumata Celtis*. Edizione critica e commento». *Eikasmós* 32 (2021): 247-272.

[2] Cammisuli, Salvatore. «La sezione sui colori nel glossario degli *Hermeneumata Celtis*: edizione critica e commento». *Wiener Studien* 134 (2021): 199-221.

[3] Cipolla, Paolo Biagio. «Su alcune glosse degli *Hermeneumata Celtis*». *Commentaria Classica. Studi di filologia greca e latina* 7 (2020): 115-135.

[4] Dickey, Eleanor, (a cura di). *The Colloquia of the Hermeneumata Pseudodositheana*, Volume 2: *Colloquium Harleianum, Colloquium Montepessulanum, Colloquium Celtis, and Fragments*. Cambridge: Cambridge University Press, 2015.

[5] Dionisotti, Anna Carlotta. «From Ausonius' schooldays? A schoolbook and its relatives». *The Journal of Roman Studies* 72 (1982): 83-125.

[6] Ferri, Rolando. «*Hermeneumata Celtis*. The making of a late-antique bilingual glossary». In *The Latin of Roman Lexicography*, (a cura di) Rolando Ferri, 141-169. Pisa – Roma: Serra, 2011.

[7] Ferri, Rolando. «Textual and linguistic notes on the *Hermeneumata Celtis* and the *Corpus glossariorum*». *Classical Quarterly*, n. s., 60 (2010): 238-242.

[8] Ferri, Rolando. «Vulgar Latin in the bilingual glossaries: the unpublished *Hermeneumata Celtis* and their contribution». In *Latin vulgaire – latin tardif IX. Actes du IX Colloque international sur le latin vulgaire et tardif*, (a cura di) Frédérique Biville, Marie-Karine Lhommé, e Daniel Vallat, 753-763. Lyon: Maison de l'orient et de la Mediterranée - Jean Pouilloux, 2012.

[9] Ferri, Rolando, e Anna Zago. «Isidoro e i vocabolari antichi dell'uso». *Archivum Latinitatis Medii Aevi* 77 (2019): 73-95.

[10] Gatti, Paolo. «Nomi di pesci negli *Hermeneumata Celtis*». *Archivum Latinitatis Medii Aevi* 64 (2006): 105-121.

[11] Kraft, Ulrich. «Περὶ χρυσέων κοσμημάτων. Ein Titulus aus dem lateinisch-griechischen Celtis-Glossar». In *Von Sklaven, Pächtern und Politikern. Beiträge zum Alltag in Ägypten, Griechenland und Rom.* Δουλικὰ ἔργα *zu Ehren von Reinhold Scholl*, (a cura di) Lutz Popko, Nadine Quenouille, e Michaela Rücker, 139-163. Berlin – New York: De Gruyter, 2012.

[12] Kramer, Johannes. «Die Ämterliste aus dem Wiener Celtis-Glossar». In *Wiener Papyri. Als Festgabe zum 60. Geburtstag von Hermann Harrauer*, (a cura di) Bernhard Palme, 249-265. Wien: Holzhausen, 2001.

[13] Kramer, Johannes. «Lateinisch-griechisches Glossar: Celtis' Abschrift aus einem Papyruskodex». In *Paramone. Editionen und aufsatze von Mitgliedern des Heidelberger Instituts fur Papyrologie zwischen 1982 und 2004*, (a cura di) James M. S. Cowey e Barbel Kramer, 43-62. München – Leipzig: K. G. Saur, 2004.

[14] Lipani, Sara. «I lemmi sulla navigazione nella sezione 47 del glossario degli *Hermeneumata Celtis*». *L'Archeologo subacqueo* 29 (2023): 1-28.

[15] Ortoleva, Vincenzo. «Gli *Hermeneumata Celtis*: osservazioni a proposito di alcuni studi recenti». *Wiener Studien* 131 (2018): 229-272.

[16] Ortoleva, Vincenzo, Maria Rosaria Petringa, e Salvatore Cammisuli (a cura di). *Onomastikón, Studi di lessicografia greca e latina. Prima edizione critica del Glossario degli* Hermeneumata Celtis. Catania, 2020. http://onomastikon.altervista.org.

[17] Petringa, Maria Rosaria. «A proposito di due passi della parafrasi del libro di *Giosuè* nel poema dell'*Heptateuchos*». *Paideia* 73 (2018): 1423-1427.

[18] Petringa, Maria Rosaria. «A proposito di una recente (parziale) edizione del *Liber Exodus* del poema dell'*Heptateuchos*». *Commentaria Classica. Studi di filologia greca e latina* 10 (2023): 373-384.

[19] Petringa, Maria Rosaria.«Adamo ed Eva e il frutto proibito nel poema dell'*Heptateuchos* (*gen*. 64-90). Testo critico, traduzione e commento». *Commentaria Classica. Studi di filologia greca e latina* 4 (2017): 105-118.

[20] Petringa, Maria Rosaria. «Alcune emendazioni inedite di Giuseppe Giusto Scaligero ai carmi pseudociprianei». *Commentaria Classica. Studi di filologia greca e latina* 1 (2014): 109-117.

[21] Petringa, Maria Rosaria. «Alcune note esegetiche di Giuseppe Giusto Scaligero al testo dei carmi pseudociprianei». *Commentaria Classica. Studi di filologia greca e latina* 2 (2015): 99-108.

[22] Petringa, Maria Rosaria. «Alcune particolarità linguistiche nell'anonimo poema dell'*Heptateuchos*». In *A Current Perspectives on Latin Grammar, Lexicon and Pragmatics. Selected Papers from the 20th International Colloquium on Latin Linguistics, Linguisticae Dissertationes. Universidad de Las Palmas de Gran Canaria, Spain, June 17-21, 2019*, (a cura di) Antonio M. Martín Rodríguez, 227-234. Madrid: Ediciones Clásicas, 2021.

[23] Petringa, Maria Rosaria. «*Christi memor, immemor aevi*. La memoria di Cristo nel carme 15 di Paolino di Nola». In *La memoria. Forme e finalità del ricordare nel cristianesimo antico, XLVIII Incontro di Studiosi dell'Antichità Cristiana.* Roma, Institutum Patristicum Augustinianum, 5-7 maggio 2022, Studia Ephemeridis Augustinianum:77-84. 164. Firenze: Nerbini International, 2023.

[24] Petringa, Maria Rosaria. «Giovenco, *Evangeliorum libri* 4,657-664». *Commentaria Classica. Studi di filologia greca e latina* 3 (2016): 113-120.

[25] Petringa, Maria Rosaria. «Il *De mortibus boum* di Endelechio». In *La Veterinaria antica e medievale. Testi greci, latini, arabi e romanzi, Atti del II Convegno internazionale*, (a cura di) V. Ortoleva e M. R. Petringa, 243-258. Lugano: Lumières Internationales, 2009.

[26] Petringa, Maria Rosaria. «Il lessico del vestiario nel glossario latino-greco degli *Hermeneumata Celtis*». In *Proceedings of the 22nd International Colloquium on Latin Linguistics*. Charles University Prague, 2023.

[27] Petringa, Maria Rosaria. «Il lessico della medicina nel glossario latino-greco degli *Hermeneumata Celtis*». In *Santé et maladie, diététique et thérapeutique, Proceedings of the XVe Colloquium international Textes médicaux latins et présalernitains*, (a cura di) J.-C. Coutil- V. Gitton. Université Toulouse, 2024.

[28] Petringa, Maria Rosaria. «Il lessico dell'agricultura nel glossario latino-greco degli *Hermeneumata Celtis*». In *Proceedings of the 15th International Colloquium on Latin Vulgaire - Latin Tardif*. München, *Thesaurus Linguae Latinae* (Bavarian Academy of Sciences and Humanities), 2024.

[29] Petringa, Maria Rosaria. «Il paradiso terrestre nella riscrittura del poeta dell'*Heptateuchos* (*gen*. 64-133): analisi del lessico delle emozioni». In *Spazi e tempi delle emozioni. Dai primi secoli all'età bizantina, Atti del Convegno "Progetto FIR 2014" e delle VI Giornate di Studio della CULCA*, 185-207. Acireale – Roma: Bonanno Editore, 2018.

[30] Petringa, Maria Rosaria. *Il poema dell'*Heptateuch*os.* Itinera philologica *tra tardoantico e alto medioevo*. (Biblioteca di *Commentaria Classica*, I). Catania: Litterae Press, 2016.

[31] Petringa, Maria Rosaria. «Il *signum crucis* nel *De mortibus boum* di Endelechio (vv. 97-132)». (a cura di) Maria Antonietta Barbàra e Maria Rosaria Petringa. *Commentaria Classica. Studi di filologia greca e latina* 6 Supplemento, *Tenax memoria*, In ricordo di Sandro Leanza (2019): 147-175.

[32] Petringa, Maria Rosaria. «L'aggettivo *innumerosus* nel poema dell'*Heptateuchos* (*exod*. 7)». *Commentaria Classica. Studi di filologia greca e latina* 8 (2021): 215-222.

[33] Petringa, Maria Rosaria. «Le attestazioni del verbo *clepto* nel latino tardo e medievale». In *Latin Vulgaire – Latin Tardif 10, Actes du Xe Colloque international sur le latin vulgaire et tardif*, (a cura di) C. Fedriani P. Molinelli P. Cuzzolin, 615- 626. Biblioteca di Linguistica e Filologia, II. Bergamo: Sestante Edizioni, 2014.

[34] Petringa, Maria Rosaria. «Lo strano caso del fr. 11 del poema dell'*Heptateuchos*: storia di incomprensioni vecchie e nuove». *Commentaria Classica. Studi di filologia greca e latina* 3 (2016): 121-127.

[35] Petringa, Maria Rosaria. «Nuovi contributi sulla lingua dell'anonimo poema dell'*Heptateuchos* (i termini *anus* e *odium*)». In *Varietate delectamur: Multifarious Approaches to Synchronic and Diachronic Variation in Latin. Selected Papers from the 14th International Colloquium on Late and Vulgar Latin, Ghent University, 5-9 September, 2022*. Turnhout, *sub prelo*.

[36] Petringa, Maria Rosaria. «Particolarità lessicali nel poema dell'*Heptateuchos*». *Commentaria Classica. Studi di filologia greca e latina* 5 (2018): 57-60.

[37] Petringa, Maria Rosaria. «Un frammento di un'anonima parafrasi metrica del *Liber Genesis* nel cod. Oxford, Bodleian Library, Canon. Bibl. Lat. 80: edizione critica, traduzione e commento». *Commentaria Classica. Studi di filologia greca e latina* 9 (2022): 197-207.

[38] Petringa, Maria Rosaria. «Un problema testuale in Endelechio». *Commentaria Classica. Studi di filologia greca e latina* 7, fasc. In memoria di Antonio Vincenzo Nazzaro (2020): 81-85.

[39] Petringa, Maria Rosaria. «Uno pseudogrecismo fortunato: a proposito della forma *haemorrhoissa* nei testi patristici». *Commentaria Classica. Studi di filologia greca e latina* 6 (2019): 9-17.

[40] Petringa, Maria Rosaria, e Paladini Mariantonietta. «Eupoli fr. 391 K.-A.: fra Giuliano e Gregorio di Nazianzo». In *Templa serena. Studi in onore di Enrico Flores*, 35-41. Napoli: FedOA Press – Federico II University Press, 2021.

The Corr<si>Ca Project: enhancing and "querying" the Canioni family correspondence

Tiziana Pasciuto[1], Selenia Anastasi[2], Daniele Zolezzi[3], Simonetta Acacia[4], Giada D'Ippolito[5], Chiara Storace[6], Maria Tolaini[7]

[1] CNR IMATI Genoa & Department of Modern Languages and Cultures, University of Genoa, Italy - tiziana.pasciuto@edu.unige.it

[2] Department of Modern Languages and Cultures, University of Genoa, Italy & Language Technology Group, Hamburg University, Germany - selenia.anastasi@edu.unige.it

[3] Department of Modern Languages and Cultures, University of Genoa, Italy - daniele.zolezzi@edu.unige.it

[4] Department of Modern Languages and Cultures, University of Genoa, Italy - simonetta.acacia@edu.unige.it

[5] Department of Modern Languages and Cultures, University of Genoa, Italy - giadadippolito30@gmail.com

[6] Department of Modern Languages and Cultures, University of Genoa, Italy - chiara.storace@edu.unige.it

[7] Department of Modern Languages and Cultures, University of Genoa, Italy - maria.tolaini@edu.unige.it

ABSTRACT[1]

The Corr<si>Ca project, undertaken by professors and PhD students from the University of Genoa's PhD program in Digital Humanities, involves the digitization and the enhancement of a family correspondence comprising 270 letters. This correspondence belongs to the Canioni family, originally from Olmi-Cappella, an inland village in the northern part of Corsica. The letters span the years 1882-1918, featuring contributions from both male and female writers with varying levels of literacy in the two languages of correspondence, namely Italian and French.

This paper outlines the ongoing development of (i) a blog designed to enrich the document corpus, engaging not only general users but also secondary school students, and (ii) an ontology for the future querying of this archival material. The ontology aims to extract valuable references and explore the correlations between language, gender, and the authors' origins. The objective is to delve into archival, philological, and sociolinguistic aspects related to the reference population.

KEYWORDS

Dissemination; Correspondence; Corsica; Blog; Ontology.

1. INTRODUCTION

The Canioni family correspondence, accessed with the permission of the family descendants, presents an interesting case of familial correspondence spanning the late 19th to early 20th centuries, specifically between 1882 and 1918. The 270 letters exchanged among Canioni family members and their entourage, primarily based in the Haute Corse village of Olmi-Cappella, reveal a linguistic and cultural transition from Italian to French. Written by three generations of Canioni individuals, including both semi-literate men and women, the letters provide linguistic insights and historical perspectives. Beyond linguistic shifts, the correspondence offers valuable details on daily life, material culture, local economy, politics, and trade between Corsica and the mainland, making it a significant source for linguistic and historical studies.

With the aim of enhancing this interesting corpus, valuable both from a linguistic and cultural point of view (since the chronological range of correspondence coincides with the end of the First World War), ongoing efforts include the creation of an accessible blog and the development of an ontology. The details will be discussed in the following paper's sections. Simultaneously, another sub-working group within the project is dedicated to the digitization of letters and transcription in XML/TEI format[2].

2. ENHANCING THE PROJECT: DISSEMINATION. Corr<si>Ca blog

In order to disseminate the project to as wide an audience as possible, have created a blog, providing a platform to share all the gathered material related to the Canioni family. Although opinions on the matter may vary, the term 'weblog' is widely credited to Jorn Barger [7]. 'Blog' is employed as an English onomatopoeic term, evoking the act of vomiting, and infusing an additional meaning to the term, depicting it as a platform where one can freely regurgitate and share their

[1] Author's statement - the author of the present paper contributed to the writing of this article as follows: T. Pasciuto (1. Introduction, 2. Corr<si>Ca blog, 3. Corr<si>Ca Ontology, 4. Conclusions and future works), S. Anastasi (3. Corr<si>Ca Ontology), D. Zolezzi (2. Corr<si>Ca blog), S. Acacia (2. Corr<si>Ca blog), G. D'Ippolito (3. Corr<si>Ca Ontology), C. Storace (2. Educational section), M. Tolaini (2. Educational section).

[2] For further information on this aspect, please refer to the contribution of Giaufret *et al.*, «Il progetto Corr<si>Ca: edizione digitale della corrispondenza Canioni», in the present volume.

content [6]. In recent years, several digital platforms emerged with the goal of enhancing cultural heritage and narrating historical and cultural events, both on a local and global scale, to raise awareness within the worldwide community. These initiatives aim to disseminate knowledge and foster appreciation for cultural richness through digital tools accessible to a broad audience.

For the Corr<si>Ca project, various platforms served as inspiration, for example the WarSampo Portal[3] [8-9], which reassembles life stories of soldiers fallen in World War II, the Italian Bellini Digital Correspondence[4] [2-3, 4, 12], the digital edition of the letters of the Sicilian composer Vincenzo Bellini, and the platform MythLod[5] [10, 11], born for the formal representation of cultural metadata.

Corr<si>Ca blog is powered by WordPress, a highly customizable Content Management System (CMS) widely used for creating and managing websites. This decision was motivated by our commitment to make the project's website as inclusive and open as possible, encouraging interaction from all visitors through the ability to comment on content. The flexibility of WordPress enables us to tailor the appearance and functionality of our site to effectively meet the needs of our community and visitors [13]. The user-friendly interface facilitates navigation and interaction, fostering the development of meaningful connections, promoting the sharing of knowledge, and nurturing a constructive dialogue to cultivate an active community.

The structure and the contents of the Corr<si>Ca blog are accessible in two languages (French and Italian), while the letters are available only in the language in which they are written, in order to respect the original language and avoid providing translations that could be artificial, as they would be unable to reproduce many features of the original such as words hypo- or hypersegmentation.

A minimalist and straightforward style has been upheld to highlight the authentic content of the platform, avoiding excessive visual distractions. The selected theme is 'Why Minimalist Blogger X', crafted to impart an elegant look to the blog. It ensures full responsiveness and optimization for search engines, making the content more attractive to search engines and easily locatable[6].

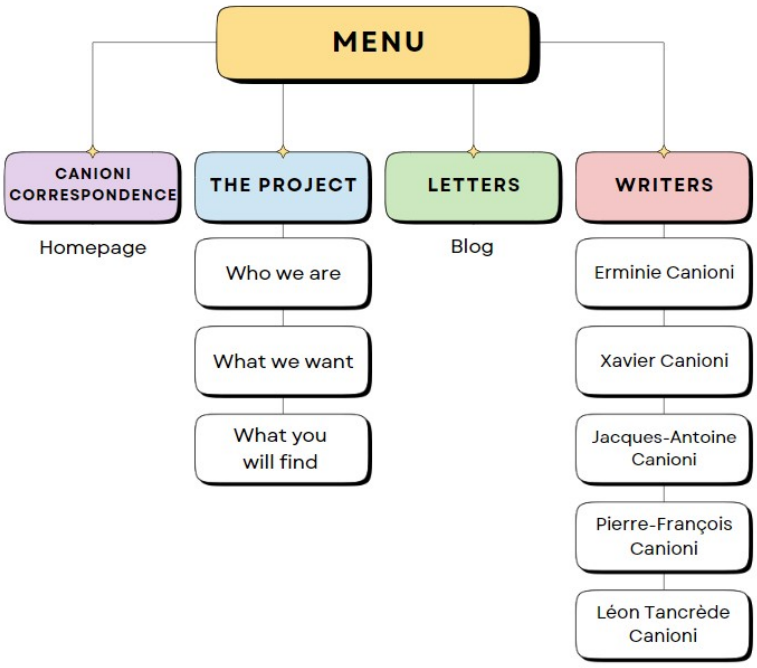


Figure 1. Graphic scheme of the main menu items of the Corr<si>Ca blog, at the present time

The goal of the project, aimed at bringing the epistolary heritage of the Canioni family back to life within the blog's virtual context, led us to structure the site to explain not only the objectives of the project and their creators, but also to provide

[3] https://www.sotasampo.fi/en/
[4] http://bellinicorrespondence.cnr.it/; http://licodemo.ilc.cnr.it/demo/bellini/build/
[5] https://dharc-org.github.io/mythlod/
[6] Minimalist-blogger_link

access to all documents in the corpus (see Fig. 1). The site supplies two different access modes for users to explore the letters' transcriptions, which are explained below.

The first option is represented by the *"Letters"* section (marked with the term *Lettere* in Italian, and *Lettres* in French), where users can peruse the transcribed letters (at the moment, only a sample, see Fig. 2a). The letters are labeled with the names of the sender and recipient, for example, *'Da Xavier a Jacques Antoine' ('From Xavier to Jacques Antoine')*, to clearly indicate to readers which family members the letter pertains to. The letters have been published on the blog by adjusting the post's publication date to the actual date on which the letter was sent (e.g., February 9, 1884), establishing a link with the present, where each letter is treated as a post on social networks. The letters feature a **diplomatic transcription of the text**, where only word segmentation, punctuation and use of capital letters have been modified to make the letters easier to read, **a scanned version** of the original document available for consultation as you progress through the reading, and **a geographical map** showing where the letter was written and received, and other places mentioned in the text. The points of interest (POI) offer additional information about the protagonists, not only pinpointing the location but also detailing their activities at that moment and specifying their degree of relationship.

Regarding the visualization of the places mentioned in the letters, in addition to the map for each document, a general map for the entire corpus is embedded in the introductory section on the Canioni correspondence. Here users can filter the places based on data deduced from the transcription, in particular fields such as author, recipient, date, type of place (origin, destination, cited), etc.

The second method for accessing information is through a search bar that can retrieve letters containing keywords entered by visitors, enabling them to quickly find content based on their interests.

The blog also includes a section featuring profiles of the main writers and recipients of the letters (see Fig. 2b). Each profile provides a brief biography and, where available, a picture of the person being discussed, allowing users to put a face to the stories and empathize with them. Users can comment on both the letters and the profiles of the authors, creating an opportunity for interaction and discussion. Furthermore, it is planned to add an educational section, which will be presented in the next paragraph.

The main objective is to build a community of practice to generate new, organized, and high-quality knowledge [16] around the themes discussed in the letters and the historical period they refer to.

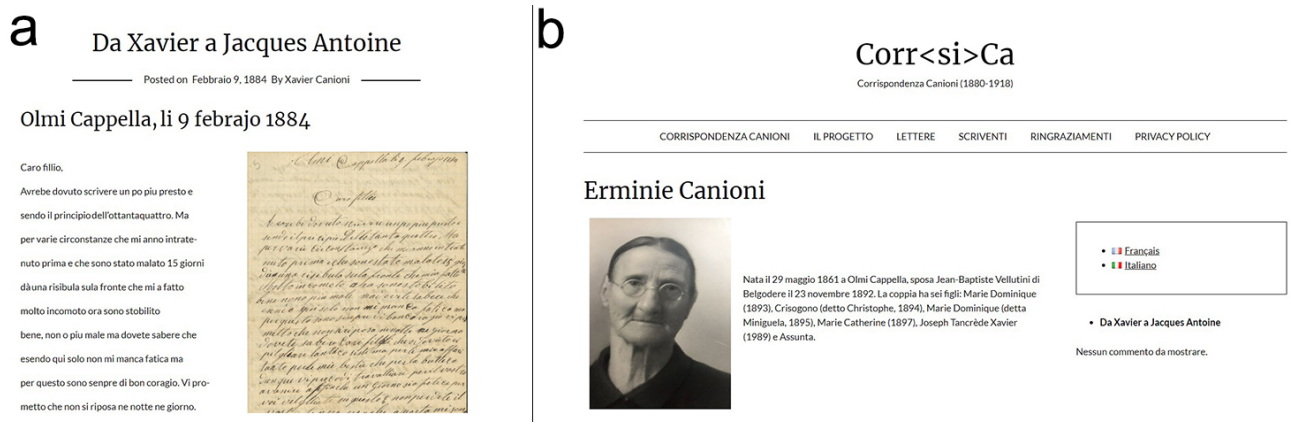


Figure 2. a) A letter of the Canioni family, featured in the "Letters" section of the site. Besides the digitalized version and the text transcription, valuable information (sender and recipient, date, and place) are provided.
b) Short biography of Erminie, one of the writers of the Canioni corpus.

Educational section

Besides offering an inclusive and engaging website interface that allows all kinds of users to enjoy the contents, it is intended to create a specific educational section for students. Thanks to the variety of topics covered, the Canioni correspondence constitutes an interesting case of epistolary exchange that can be used for linguistic, socio-cultural, economic, and historical studies. For these reasons, we believe that not only experts and enthusiasts can benefit from the dissemination of these letters but also secondary school students. The information reported by the numerous letters offers insights on different school subjects, such as Foreign Language Learning, History, Literature, and Active Citizenship. We present a proposal for the educational section to be included on the website.

To formulate it, we carried out a short analysis of existing sites dedicated to digitizing and disseminating correspondence[7]. For this purpose, we defined a set of observation criteria, starting with the main question: is there a section dedicated to education? If it was present, we also observed the focus of this section, the target audience chosen, the interactivity of the web page, what resources were made available both in terms of materials provided, exercises or teaching activities proposed and, finally, the strengths and weaknesses of the section.

This brief analysis showed that only two of these websites had a dedicated educational section, thus highlighting the originality and the need for more such contributions. In particular, the Pirandello Nazionale Website, a digital project aimed to publish the integral corpus of Pirandello's works on-line, presents a captivating educational section. It provides digital resources that include educational paths characterized by mixed media materials whose exploration is facilitated by a well-designed legend. Given the brief analysis results, and considering the highlighted valuable examples, we decided to create a rich and multimedia educational section illustrated by an index page that shows the main focuses of the Canioni Project. Each focus contains two subsections: one with in-depth materials and one with suggested educational activities.

Regarding the in-depth materials, our proposal includes open-source links for further exploration of the themes, as well as videos and audio materials. These materials allow students to explore the topics independently and enable teachers to enhance and prepare their lessons. Multimedia elements are designed to meet the diverse educational needs of students.

Concerning the teaching activities, Canioni correspondence offers various points of historical and sociological interest that could be addressed by students through the analysis of the letters. For example, a design for the teaching unit could be to guide the exploration and the in-depth study of topics discussed in various letters across different historical periods, such as weights and measures, agricultural practices, epidemics, and medical treatments. The teaching unit will consist of two parts. Initially, introductory materials will be provided to contextualize the specific themes. Teachers will use these materials to design classroom activities or to guide students in their autonomous use of the blog. The second segment of the teaching unit involves learners using keywords in the blog's search bar to retrieve information on the chosen topics. Next, students will create PowerPoint presentations to expose their understanding of the evolution of these topics.

The value of the Canioni correspondence does not end in its being historical evidence of the cultural and linguistic transition but also encompasses its being an interdisciplinary starting point for reflections at the scholastic level, at the academic one.

3. QUERYING THE PROJECT: THE CORR<SI>CA ONTOLOGY

A crucial part of the Corr<si>Ca project involves the deep encoding of the texts according to the XML/TEI standard, followed by the representation and organization of data and metadata within an ontological model, named **OntoCorr<si>Ca** (see Fig. 3). This formalization aims to enable complex queries on the database from the perspective of interactions among the actors involved in the exchanges, also extracting useful references to the places where these exchanges occurred, and the relationship between language, gender, and authors' origin, for those who are interested in investigating the archival, philological, and sociolinguistic aspects of the population. This section aims to describe, at a high level, the ontological schema created specifically for the management of the information related to the corpus. Future works dedicated to modeling sociolinguistics characteristics such as the development of written proficiency first in Italian, then in French, will also be discussed within a multilingual repertoire context where the Corsican language was the only spoken one utilized. This section of the project represents an effort in line with the European enterprise of digitizing the rich cultural heritage of the UE[8], and the subsequent creation of interoperable web resources that may provide in the future an in-depth understanding of the relationships between places, cultures, and languages whose history and tradition are deeply connected.

An initial phase of the project was dedicated to the study and selection of foundational ontological resources existing for the modeling of the archival domain and, in particular, for the representation of epistolary correspondence. In recent years, proposals concerning the semantic organization of cultural heritage have proliferated, many of which are based on the CIDOC-CRM model (Conceptual Reference Model)[9] [5], which has emerged as an international standard since 2006 for the controlled exchange of information on cultural heritage. Based on CIDOC-CRM, many ontological models have been developed to improve the semantic expressiveness of the typical features of epistolary correspondence and address specific issues previously overlooked. Nevertheless, we believe that this standard is not without problems concerning the terminology used and concepts, such as the conceptual opacity of some classes as discussed in [14-15] and the indiscriminate use of the over-extended masculine for classes such as E-22 – Man-Made Object, E24 – Physical Man-Made

[7] MythLOD; Bellini Digital Correspondence; WarSampo; Visual Correspondence; Claviuson on the web; Digitising experiences of migration; Early Modern Letters Online; Mapping the Republic of Letters; Storia digitale UniCA. Il portale della Storia Digitale dell'Università degli Studi di Cagliari; Transcribe Bentham; Edizione Nazionale dell'Opera Omnia di Luigi Pirandello.

[8] https://www.europarl.europa.eu/factsheets/en/sheet/64/l-agenda-digitale-europea/

[9] https://www.cidoc-crm.org/

Thing, E25 – Man-Made Feature, E71 – Man-Made Thing[10]. For this reason, it has been deliberately chosen to operate, at least in this initial phase, on a parallel path, without neglecting the purposes of alignment with other exemplary Italian and international models such as OntoBelliniLetters[11], MythLod, and WarSampo.
In the early stages of OntoCorr<si>Ca modeling, we explored other ontologies describing archival and bibliographic domains: this analysis will lead to the alignment of classes and properties between ontologies later on. Specifically, we examined bibliographic standards such as Functional Requirements for Bibliographic Records (FRBR)[12] and Bibliographic Framework (BIBFRAME)[13], focusing on distinguishing between paper and digital documents, copies, and drafts. Additionally, since the heritage in question consists of a corpus of letters related to archival heritage, the ontology was modeled after the Records in Contexts Ontology (RiC-O)[14], which has been available in a stable first version since December 2023.

XML-TEI elements	OntoCorr<si>Ca classes and properties
<body>	[C] LetterBody
<opener>	[subC] Opener
<closer>	[subC] Closer
<salute>	[subC] Salute
<postscript>	[subC] PostScript
<signed>	[subC] Signature
<date>	[C] ChronologicalDate
<placeName>	[C] Place
<persName>	[C] foaf:Person - [DP] hasName

Table 1. Example of mapping between XML/TEI elements (selected for labelling the transcriptions of Canioni letters corpus) and some of the classes [C], subclasses [subC] and datatype properties [DP] of the OntoCorr<si>Ca ontology.

To begin outlining the ontology classes and properties, XML/TEI elements, used to label transcriptions, have been considered (Table 1). For the modeling of data related to the corpus, the material has been represented according to three different conceptual levels mirroring possible planes of analysis.
The **first level** is the analysis of the **agents** (*foaf:Agent, foaf:Person*) and **places** (*Place*) involved in the interactions, with a particular focus on one hand on *name, gender*, and *birth* and *death dates* of the authors, and on the other hand the *source* and *destination places* of the letters.
A **second level** of analysis is aimed at capturing the internal organizational **structures of texts** in relation to the archival domain. Every *Document* is written in one or more languages (*Language*), has a unique *Identifier*, is characterized by one or more types of medium (*Matter*) and by a specific state of conservation (*StateOfConservation*). On this second level, an important ontological distinction has been made to distinguish the *Form* of the document, that can be *Paper* or *Digital,* and the *Status* of realization, such as a *Draft*, a *Copy* or a *Final* document, handwritten or not; the last three classes are in a *disjunctive* relationship. Each letter (*Letter*), which is a *type* of document (*hasType*) may have one or more *authors* (that often – but not always – in letters coincide with the *sender*) and *recipients*, who in turn may contain references to other entities such as people, organizations, places, works, events, and temporal ranges. A subdivision has been made considering the internal organization of the text (*isStructuredIn*): in fact, each letter is annotated according to the XML/TEI standard, taking into account general structural elements of the body of the text (*LetterBody*), such as opening (*Opener*) and closing phrases (*Closer*), postscripts *(PostScript*), author's signatures (*Signature*), and greetings (*Salute*). Other elements of the body of the text that belong to the original copy and that are also intended to be captured through the digital version are the presence of deletions, corrections, ambiguous or illegible graphic elements, marginal notes, and the subdivision of the text into paragraphs.
The final but important **third level** of modeling concerns the **textual and linguistic information**, with particular reference to the terminological, morphological, syntactic, and grammatical characteristics.

[10] For the latest version of the classes mentioned in CIDOC-CRM: https://www.cidoc-crm.org/Entity/e24-physical-man-made-thing/version-6.2.1.
[11] http://bellininrete.istc.cnr.it/OntoBelliniLetters.html/
[12] https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf
[13] https://www.loc.gov/bibframe/
[14] https://github.com/ICA-EGAD/RiC-O/tree/master/ontology/current-version/HTML_view

Regarding the ontological editor, the first part of the work was developed on Protégé, while for the linguistic domain the possibility of integrating the Tedi software[15], designed for the construction of multilingual ontoterminologies, will be evaluated. Tedi allows exporting ontoterminologies in RDF/OWL format, permitting the importation of the ontoterminology into Protégé.

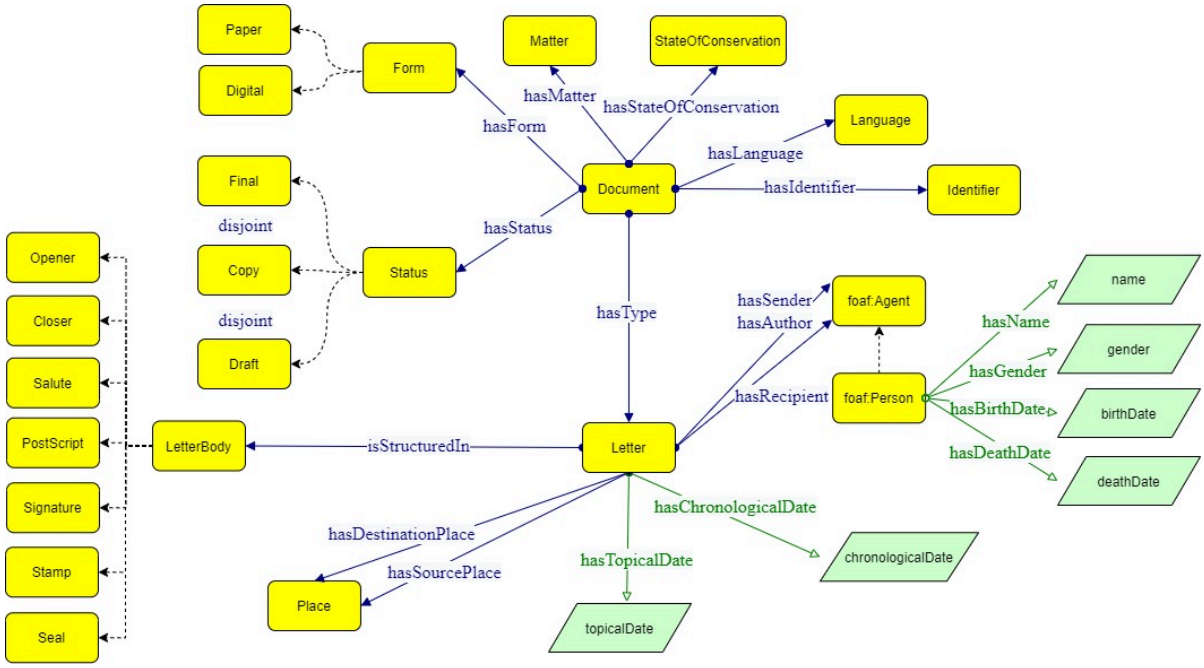


Figure 3. Graphical representation of the ontological model OntoCorr<si>Ca. The yellow boxes and green parallelograms indicate respectively the classes and the datatypes of the ontology, while the blue and green lines connecting classes and datatypes represent respectively object and datatype properties.

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we have outlined the choices and methodologies employed in creating a blog dedicated to the study and enhancement of the archival material of the Canioni family's correspondence. Additionally, we have illustrated the preliminary formalization of an ontological model designed for the querying of data.

The project is currently ongoing, and every initiative illustrated in the paper is at a prototype level. Letters and transcriptions, customizing the enjoyment of the heritage for both general users, secondary school students and possibly university students, will further enrich the blog. In addition, the ontological module will be extended with more detailed definitions related to the physical and digital versions of the letters. Finally, we plan to validate the model by instantiating the data provided by the project and enabling semantic querying of the corpus.

For the representation of linguistic features, Ontolex-Lemon will be considered. Ontolex-Lemon is a standard model for the multilingual, lexical, syntactic, and semantic modeling of terms and allows, for example, the decomposition of compound nouns, highlighting variations considering contextual meanings, and the use of metadata that provide descriptive elements on how, why, by whom, etc. [1]. Note that utilizing this ontology requires a manual intervention in transcribing letters' content into the XML/TEI format: specifically, it involves the targeted labelling of lemmas using the `<choice>` TEI element. Consequently, this process will probably be executed on a restricted set of letters, chosen for their linguistic significance.

5. ACKNOWLEDGEMENTS

We express our gratitude to the descendants of the Canioni family, who graciously provided us with access to the original correspondence that served as the focal point of this study, and to Santu Massiani, precious guardian of local history.

REFERENCES

[1] Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. *Linguistic Linked Data: Representation, Generation and Applications*. New York Inc: Springer, 2020. https://doi.org/10.1007/978-3-030-30225-2.

[15] http://ontoterminology.com/tedi/

[2] Del Grosso, Angelo Mario, Erica Capizzi, Salvatore Cristofaro, Maria R. De Luca, Emiliano Giovannetti, Simone Marchi, Graziella Seminara, and Daria Spampinato. 'Bellini's Correspondence: A Digital Scholarly Edition for a Multimedia Museum'. *Umanistica Digitale* 3, no. 7 (2019): 23-47. https://doi.org/10.6092/issn.2532-8816/9162.

[3] Del Grosso, Angelo Mario, and Daria Spampinato, eds. *Bellini Digital Correspondence*. CNR Edizioni, 2023. http://bellinicorrespondence.cnr.it.

[4] Del Grosso, Angelo Mario, Daria Spampinato, Erica Capizzi, Salvatore Cristofaro, and Graziella Seminara. 'Promoting Bellini's Legacy and the Italian Opera by Scholarly Digital Editing His Own Correspondence'. In *What Is Text, Really? TEI and Beyond*. Graz, Austria, 2019.

[5] Doerr, Martin, Christian-Emil Ore, and Stephen Stead. 'The CIDOC Conceptual Reference Model - A New Standard for Knowledge Sharing'. *Proceedings of the Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modelling, Auckland, New Zealand* 83 (5 November 2007): 51-56. https://doi.org/ 10.13140/2.1.1420.6400.

[6] Garden, Mary. 'Defining Blog: A Fool's Errand or a Necessary Undertaking'. *Journalism* 13, no. 4 (2012): 483-499. https://doi.org/ 10.1177/1464884911421700.

[7] Hookway, Nicholas. 'Entering the Blogosphere': Some Strategies for Using Blogs in Social Research'. *Qualitative Research* 8, no. 1 (2008): 91-113. https://doi.org/10.1177/1468794107085298.

[8] Hyvönen, Eero. 'Digital Humanities on the Semantic Web: Sampo Model and Portal Series'. Edited by Christoph Schlieder. *Semantic Web* 14, no. 4 (April 2023): 729-744. https://doi.org/10.3233/SW-223034.

[9] Koho, Mikko, Esko Ikkala, Petri Leskinen, Minna Tampe, Jouni Tuominen, and Eero Hyvönen. 'WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data'. Edited by Christoph Schlieder. *Semantic Web* 12, no. 2 (January 2021): 265-278. https://doi.org/10.3233/SW-200392.

[10] Pasqual, Valentina, and Francesca Tomasi. 'Data Narratives with Linked Open Data, the Case of MythLOD Storytelling'. In *Digital Humanities 2023: Book of Abstracts*, edited by Walter Scholger, Georg Vogeler, Toma Tasovac, Anne Baillot, and Paul Helling, 283-284. Graz, Austria, 2023.

[11] Pasqual, Valentina, and Francesca Tomasi, 'Linked Open Data per La Valorizzazione Di Collezioni Culturali: Il Dataset MythLOD.' *AIB Studi* 62, no. 1 (May 2022): 149-168. https://doi.org/10.2426/aibstudi-13301.

[12] Pellino, Santa, Pietro Sichera, Angelo Mario Del Grosso, and Daria Spampinato. 'Dalla Codifica Alla Fruizione: L'edizione Digitale Bellini Digital. AIUCD 2022, Lecce 1-3 Giugno 2022 Correspondence'. In *Culture Digitali. Intersezioni: Filosofia, Arti, Media. AIUCD2022*, edited by Fabio Ciracì, Giulia Miglietta, and Carola Gatto, 163-168. Lecce, 2022.

[13] Sabin-Wilson, Lisa, and Matt Mullenweg. *WordPress for Dummies*. John Wiley & Sons, 2011.

[14] Sanfilippo, Emilio M. 'Ontologies for Information Entities: State of the Art and Open Challenges'. *Applied Ontology* 16, no. 2 (2021): 111-135.

[15] Sanfilippo, Emilio M., Béatrice Markhoff, and Pittet Perrine. 'Ontological Analysis and Modularization of CIDOC-CRM'. In *Information Systems: Proceedings of the 11th International Conference (FOIS 2020)*, edited by Boyan Brodaric and Fabian Neuhaus, 107-121. IOS Press, 2020.

[16] Wenger, Etienne. *Comunità Di Pratica. Apprendimento, Significato e Identità*. Milano: Raffaello Cortina Editore, 2007.

XML/TEI e dizionari *born-digital*: una proposta per le risorse lessicografiche della rete LexiCad/Pluto

Giuseppe Leonardo Zappalà
Università di Catania, Italia – giuseppe.zappala1@unict.it

ABSTRACT

Nell'ambito del progetto *PRIN QM (Quattrocento Meridionale) – The Future of Old Italian. Towards a New Digital Lexicography with the Southern Texts Corpus*, il cui obiettivo più ambizioso è quello di creare una rete di vocabolari in collegamento dinamico all'interno del sistema di gestione LexiCad/Pluto, il presente progetto mira a definire un modello di codifica in XML/TEI per le voci del *TLIO*, applicabile a tutte le risorse lessicografiche connesse.

A partire dallo studio della microstruttura delle voci del *TLIO* e dalla codifica non standard in XML prodotta durante i lavori per il progetto ReddiX, si fornisce un primo modello di conversione dell'XML delle voci del *TLIO* in uno standard XML/TEI, integrato con i moduli forniti da TEI Lex-0, con il vantaggio di costruire un modello standard che possa essere riadattato e riutilizzato in progetti differenti, garantendo una piena interoperabilità, lo scambio reciproco di dati e la convergenza di metodi di trattamento condivisi. La codifica qui proposta sarà il punto di partenza per fornire ulteriori strumenti nel più ampio piano di integrazione fra le diverse imprese lessicografiche per l'italiano antico, con conseguente sviluppo di applicativi in grado di implementare le potenzialità della piattaforma in termini di interoperabilità e usabilità.

PAROLE CHIAVE

TEI; TLIO; TEI Lex-0; dictionary encoding; LexiCad/Pluto.

1. OBIETTIVI

Un dizionario elettronico ha come primo lettore il calcolatore, incapace di interpretare informazioni ambigue, per cui la forma utilizzata diventa un valore imprescindibile e appare essenziale la definizione di uno schema di codifica che riesca a rappresentare la struttura della voce secondo un modello coerente e uniforme.

All'interno delle risorse informatiche sviluppate per il *Tesoro della Lingua Italiana delle Origini* (*TLIO*), il progetto ReddiX [6, 7] – avviato nel 2011 per lo sviluppo di un *Dictionary Writing System* (*DWS*) che andasse a sostituire la redazione delle voci in Word – pur non essendo entrato in funzione nella sua interezza, ha avuto il merito di definire una *Document Type Definition* (*DTD*) per la redazione e la pubblicazione *on-line* delle voci gestite dal sistema TLIOWeb. ReddiX ha, così, portato all'estrazione delle voci da una versione in Word a una versione codificata in XML, che scompone la voce in piccoli segmenti che ne definiscono la struttura. Il successivo passo è stato quello di pensare a una piattaforma *on-line* unica, qual è Pluto (Piattaforma Lessicografica Unica del Tesoro delle Origini) [3], che ha «imposto un nuovo paradigma di *DWS* per gli antichi volgari italiani» [1: 78], mettendo alla prova l'ipotesi centrale della sperimentazione di LexiCad[1] [1], ovvero la possibilità di separare un livello lessicografico più astratto, trasferibile integralmente su altre piattaforme, dalle personalizzazioni richieste per il *TLIO*. Nella progettazione di Pluto, pur partendo dalla *dictionary grammar* [12: 784] del *TLIO* ricostituita dal progetto ReddiX, consapevoli di alcune incongruenze nella descrizione della microstruttura, si è deciso di operare «con un approccio 'descrittivo' più che 'normativo'» [2: 106-107] per l'analisi sui *file* XML.

Nell'ambito del progetto PRIN *QM (Quattrocento Meridionale) – The Future of Old Italian. Towards a New Digital Lexicography with the Southern Texts Corpus*[2], il cui obiettivo più ambizioso è quello di creare una rete di vocabolari in collegamento dinamico all'interno del sistema di gestione LexiCad/Pluto[3], il presente progetto mira a definire un modello di codifica in XML/TEI della voce del *TLIO*, applicabile alle diverse risorse lessicografiche connesse. Il modello di partenza è quello elaborato dall'Opera del Vocabolario Italiano (OVI) per il *TLIO*, impianto lessicografico di alto valore scientifico e istituzionale per la lessicografia italoromanza, oltre che vocabolario elettronico in senso forte in virtù del modello che prevede tutti i marcatori strutturali richiesti. Le diverse imprese lessicografiche nate a partire dal *TLIO* hanno

[1] Per una presentazione del sistema si rimanda al contributo di Salvatore Arcidiacono e Antonella Zammataro in questo stesso volume.
[2] P.I del PRIN è Per Gunnar Lärson per il Consiglio Nazionale delle Ricerche, mentre Nicola De Blasi è coordinatore per l'Università degli studi di Napoli "Federico II" e Salvatore Arcidiacono per l'Università degli studi di Catania.
[3] Per limitarci alle sole risorse lessicografiche gestite attraverso LexiCad/Pluto, accanto al *TLIO*, si segnalano l'*Atlante Grammaticale della Lingua Italiana delle Origini* (*AGLIO*), il *Vocabolario Dantesco* (*VD*), il *Vocabolario Dantesco Latino* (*VDL*), il *Vocabolario storico-etimologico del veneziano* (*VEV*), il *Vocabolario del Siciliano Medievale* (*VSM*), e il *Dizionario Etimologico e Storico del Napoletano* (*DESN*) di prossima implementazione.

previsto alcuni scostamenti dal modello, pur mantenendo l'impianto lessicografico generale, così da permettere la definizione di una codifica strutturale di base che riesca a essere fruttuosamente impiegata.
La formulazione di un primo modello XML/TEI per uno dei progetti nati a partire dal *TLIO*, ovvero il *Vocabolario del Siciliano Medievale* (*VSM*)[4], aveva avuto l'obiettivo di dotare lo stesso di un livello di astrazione sul quale costruire i dati e le loro relazioni [2: 11-12] a partire dallo standard TEI, scelto per le potenti e flessibili caratteristiche della marcatura e per il prestigio accademico acquisito che, nel corso degli anni, lo ha reso lo standard *de facto* per chiunque si occupi di informatica umanistica.
La conversione dall'attuale codifica XML in una codifica standard TEI, integrata con quanto previsto dal modulo TEI Lex-0, mira a fornire ulteriori strumenti nel più ampio piano di integrazione fra le diverse imprese lessicografiche sull'italiano antico, con conseguente sviluppo di applicativi in grado di implementare le potenzialità della piattaforma in termini di interoperabilità e usabilità.

2. CODIFICA XML PER I VOCABOLARI

L'*Institute for Corpus Linguistics and Text Technology* (ICLTT), nell'ambito dei progetti lessicografici portati avanti per la gestione di dati lessicografici digitalizzati e digitali, ha sviluppato un modello di codifica basato sulle linee guida TEI che, nella versione P5, hanno ampliato la prospettiva di codifica verso qualunque «computational lexica and similiar resources for use by language-processing software» [9: 247], comprendendo così le risorse di lessicografia elettronica [9] e superando il forte orientamento verso risorse digitalizzate della versione precedente. L'ICLTT ha fornito uno schema (*ICLTT's TEI Schema*) utilizzato con successo per la codifica dei dati lessicografici all'interno dell'Istituto e pensato come sistema multiuso rivolto sia agli utenti sia alle applicazioni software. Ulteriori passi in avanti nella costruzione di uno schema condiviso per le risorse lessicografiche sono stati compiuti nell'ambito del progetto COST *Action European Network of e-Lexicography* (ENeL)[5], che ha tentato di fornire un approccio comune alla lessicografia elettronica attraverso standard riconosciuti e soluzioni condivise. Al suo interno è nata la prima bozza del progetto TEI Lex-0, poi proseguito dal gruppo di lavoro DARIAH *Lexical Resources*[6] e oggi supportato dall'infrastruttura lessicografica europea ELEXIS[7]. TEI Lex-0 fornisce un set di raccomandazioni per la codifica di dizionari *machine-readable*, partendo dalle linee guida fornite da TEI, con particolare riferimento al modulo *9 Dictionaries*, e fornendo una personalizzazione dello stesso schema TEI per facilitare l'interoperabilità tra risorse lessicali codificate in modo eterogeneo. Il progetto non mira a sostituire il capitolo dedicato nelle *Guidelines* TEI[8], ma a fornire uno strumento attraverso cui i dizionari TEI esistenti possano essere trasformati per essere interrogati, visualizzati o estratti in modo uniforme, e con cui le nuove risorse lessicografiche possano operare attraverso un *subset* più ristretto e con maggiori vincoli. I moduli offerti da TEI Lex-0[9] forniscono, infatti, una rappresentazione più precisa di alcuni fenomeni lessicali per mezzo di attributi e valori ben codificati e strutturati specificamente per le risorse lessicografiche, riducendo la molteplicità di soluzioni per la codifica di uno stesso fenomeno prevista da TEI, già messa in evidenza dal gruppo del *Lexical Markup Framework*[10] [16, 17].
Le voci di un dizionario sono oggetti altamente strutturati, costituenti nel loro insieme la macrostruttura del dizionario, che forniscono una specifica informazione riguardo all'oggetto descritto, secondo una struttura e un ordinamento fisso [13: 1195], che può essere rappresentato come «una struttura ricorsiva composta, ad ogni livello, da uno o più nodi» [11: 113]. Pur nell'ambiguità e nella varietà delle scelte possibili, l'architettura della voce deve sempre essere evidente, così che l'utente – e il calcolatore – possa decodificare le strutture lessicografiche: «each entry must be, so to say, self-contained» [20: 16].
Nel dizionario cartaceo, saranno le porzioni della microstruttura, e i relativi espedienti tipografici associati, a rendere evidente al lettore il confine tra gli elementi della voce, per cui nella codifica informatica tali elementi andranno marcati

[4] Si tratta di un vocabolario diacronico *online* del siciliano (sec. XIII seconda metà – XVI prima metà), fondato sullo spoglio del corpus ARTESIA (Archivio Testuale del Siciliano Antico), diretto da Mario Pagano. In rete: http://artesia.unict.it/vocabolario.
[5] https://www.eurac.edu/en/institutes-centers/institute-for-applied-linguistics/projects/enel
[6] https://www.dariah.eu/activities/working-groups/lexical-resources/
[7] https://elex.is/
[8] TEI Consortium (eds.). "9 Dictionaries." TEI P5: Guidelines for Electronic Text Encoding and Interchange. [4.7.0]. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.
[9] Tasovac, Toma, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, Alexander Geyken, Axel Herold, Vera Hildenbrandt, Mohamed Khemakhem, Boris Lehečka, Snežana Petrović, Ana Salgado and Andreas Witt. *TEI Lex-0: A baseline encoding for lexicographic data*. Version 0.9.2. DARIAH Working Group on Lexical Resources, 2018. https://dariah eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.
[10] Prodotto all'interno del gruppo di lavoro ISO/TC 37 (codice ISO-24613-1:2024), è uno standard ISO per il *Natural Language Processing* (NLP) e per i dizionari *machine-readable* (MRD) con lo scopo di standardizzare principi e metodi relativi alle risorse linguistiche nei contesti della comunicazione multilingue.

con una dichiarazione esplicita e univoca, tale da permettere una corretta gestione e interpretazione di ciascun blocco informativo. All'interno dell'architettura di un dizionario elettronico, con la trasformazione in *Machine-Readable Form* (MRF), la microstruttura diviene infatti l'aspetto più critico del processo ed è, dunque, il momento iniziale nella costruzione del dizionario elettronico.
Analizzata la microstruttura prevista per le voci del *TLIO* e confrontata con quella delle altre risorse lessicografiche, si è deciso di partire dal modello in una codifica non standard in XML previsto dal progetto ReddiX, così da operare una conversione senza perdita di dati, fornendo ulteriori specifiche nel modello per fenomeni strutturali e linguistici non precedentemente valutati. È stata effettuata un'analisi preventiva delle diverse entrate lessicali presenti nella piattaforma, giungendo alla conclusione che si tratta sempre di unità a sé stanti, registrate secondo un codice numerico univoco, anche nel caso di voci di composti e derivati o di voci di omonimi. Un caso specifico è rappresentato dalle voci di solo rinvio, per le quali è prevista la sola entrata con collegamento alla voce di base, pur rimanendo valida l'univocità della voce in termini identificativi.
Le voci verranno codificate singolarmente per permettere una piena interoperabilità delle risorse e la possibilità di collegare i diversi vocabolari attraverso un attributo di rimando alla corrispettiva voce del *TLIO*, dizionario lemmatizzato e punto di unione delle diverse risorse lessicografiche gestite dal sistema LexiCad/Pluto. Ciò permetterà di risolvere possibili sovrapposizioni e casi di lemmi omografi fra le voci e di archiviare le diverse revisioni della codifica in *file* XML separati. Il nome del *file* generato sarà composto dalla sigla del dizionario, seguito dal numero identificativo della voce corrispondente.
La codifica in XML/TEI non riguarderà la redazione, e quindi non andrà a influire su un flusso redazionale già avviato e consolidato nelle diverse esperienze lessicografiche in modalità differenti; verrà, altresì, applicata a posteriori e in forma automatica alle voci, attraverso uno *script* che permetterà il salvataggio e l'esportazione delle stesse in formato XML/TEI, secondo le modalità stabilite nel modello.

3. MODELLO DI CODIFICA

Nonostante sia indubbia l'eterogeneità dell'annotazione linguistica, con conseguente difficoltà nell'interoperabilità e nella riusabilità delle risorse linguistiche, il progetto dell'ICLTT [9], la proposta di interoperabilità fra LMF e TEI [17] e le personalizzazioni portate avanti da TEI Lex-0 [4, 8, 19] mostrano la volontà di creare una fitta rete di risorse linguistiche e lessicografiche aderenti allo standard messo a disposizione da TEI e coerente con il settore lessicografico. Il modello qui proposto (Tab. 1) intende inserirsi nell'alveo di questi progetti, nella volontà di proporre uno schema unico per le risorse lessicografiche dei volgari italiani, idealmente applicabile a nuovi progetti.
Ogni specifica sezione della microstruttura è stata codificata all'interno dell'elemento `<dictScrap>`, che racchiude parte di una voce del dizionario e che serve, nel caso specifico, a codificare separatamente le sezioni della microstruttura, riuscendo a operare in modo coerente e interoperabile fra le diverse imprese lessicografiche connesse, anche nei casi di difformità fra le diverse sezioni. Ogni elemento `<dictScrap>` sarà così costituito da un attributo @xml:id con valore costituito da "SIGLADIZIONARIO.numerovoce.siglasezione".
Rispetto a quanto proposto da LMF e da TEI, tenendo in considerazione la personalizzazione di TEI Lex-0, ci si muoverà secondo lo schema seguente (Tab. 1) per le categorie principali della voce.

LMF	**TEI**	**Modello (con integrazioni TEI Lex-0)**
LexicalEntry	`<entry type= "abbr" "affix" "foreign" "hom" "main" "supplemental" "xref">`	`<entry type= "homonymicEntry" "mainEntry" "relatedEntry", "wordFamily">`
Lemma	`<form type= "lemma">`	`<form type= "lemma">`
Word Form	`<form type= "variant">`	`<form type= "variant">`
writtenForm	`<orth>`	`<orth>`
partOfSpeech	`<pos>`	`<gram type="pos">`
grammaticalNumber	`<number>`	`<gram type="number">`

Tabella 1. Lo schema qui riprodotto prende spunto e amplia quello proposto da Romary [18: 56].

Le sezioni attualmente già a compilazione automatica, come il punto 0.3 della voce per la prima attestazione e il punto 0.5 per le poliremiatiche, continueranno a essere gestite automaticamente all'interno dei rispettivi elementi di codifica previsti attraverso alcuni *script* in via di sviluppo. Per le altre sezioni, si presentano di seguito alcuni casi esemplificativi.

- <entry>

<entry> costituisce l'elemento primario della codifica della voce e racchiude al suo interno l'attributo identificativo per il lemma (@xml:id="SIGLADIZIONARIO.numerovoce.lemma"), l'attributo identificativo per la lingua (@xml:lang) in accordo con l'ISO Standard 639-2 e l'attributo @type per l'identificazione della tipologia di voce, con valori mutuati dallo schema TEI Lex-0 ("mainEntry"; "homonymicEntry"; "relatedEntry", "wordFamily"). Nel caso di entrate omografe e di rimando, si utilizzeranno gli specifici valori adottati da TEI Lex-0 all'interno dell'attributo @type, rispettivamente il valore "relatedEntry" per le voci di rimando con identificativo della voce di rimando all'interno dell'attributo @sameAs, e valore "homonymicEntry" per le voci omonime con identificativo della voce da disambiguare all'interno dell'attributo @exclude (vd. Fig. 1). Tale soluzione permette di uniformare le scelte operate da TEI e dal modello TEI Lex-0 e di alleggerire la codifica della voce, escludendo possibili *tag* come <hom> previsti da TEI e non coerenti all'interno di un vocabolario che prevede entrate separate e dal valore identificativo esclusivo. In altre parole, questa soluzione ci svincola dalla classificazione delle voci omonime e di rimando all'interno di un'unica entrata.

```
<entry xml:id="TLIO.numerovoce.lemma" xml:lang="ita" type="mainEntry">
   <!-- Voce di rimando: <entry xml:id="TLIO.numerovoce.lemma" xml:lang="it" type="relatedEntry" sameAs="TLIO.numerovoce.lemmadirimando"> -->
   <!-- Voce omonima: <entry xml:id="TLIO.numerovoce.lemma" xml:lang="it" type="homonymicEntry" exclude="#TLIO.numerovoce.lemmaomonimo"> -->
```

Figura 1. <entry>

La forma da accogliere a lemma viene codificata all'interno dell'elemento <form> con attributo @type e valore "lemma" in posizione isolata e primaria rispetto ai diversi <dictScrap> che codificano la microstruttura della voce (vd. Fig. 1). Si decide di separare la forma accolta a lemma dalle forme attestate, che costituiscono nel loro insieme il formario – codificato nel punto 0.1 della voce con il valore "variant" dell'attributo @type – data la natura convenzionale del lemma, che «definisce un'entità virtuale, non un dato» [5: 240].

- <gramGrp>

Assecondando quanto già indicato da Romary, per superare le ambiguità previste da TEI [17: 58] e con riferimento al modello TEI Lex-0, si è deciso di codificare le informazioni grammaticali in modo più specifico, differenziando le informazioni per genere, numero, categoria ed, eventualmente, altre eventuali indicazioni tipologiche, come la valenza o la transitività del verbo, all'interno dell'elemento <gram> con specifici valori dell'attributo @type. (vd. Fig. 2).

```
<form type="lemma">
   <orth></orth></form>
<gramGrp>
   <gram type="pos"></gram>
   <gram type="gender"></gram>
   <gram type="number"></gram>
   <gram type="valency"></gram>
</gramGrp>
```

Figura 2. <gramGrp>

- <etym>

L'elemento <etym> è stato ridefinito a partire dal modulo offerto da TEI Lex-0, attraverso un insieme più specifico di opzioni per codificare ogni singolo dato etimologico [8].
La sezione relativa all'etimo, prevista nel punto 0.2 del *TLIO* e degli altri dizionari presi in considerazione, pone alcuni problemi di codifica, data la sua doppia natura. È, infatti, previsto che l'etimo venga indicato all'interno di un campo a testo libero, così da permettere a chi redige la voce di fornire sia le informazioni di base sull'etimo, con la citazione del lessico etimologico di riferimento, sia eventuali rimandi a ulteriori studi con ipotesi etimologiche aggiuntive. La forma base dell'etimo, senza rimando alla risorsa etimologica, viene invece associata alla voce, attraverso una specifica tabella gestita all'interno del *database* con associazione molti-a-molti. Si è dunque deciso di codificare le informazioni etimologiche all'interno dell'apposito *tag* <etym>, differenziando gli elementi non strutturati con l'elemento <seg>, attributo @type e valore "desc", dagli elementi già strutturati. In particolare, si prevederà un elemento <cit> con attributo @type e valore "etymon", all'interno del quale verrà codificata la lingua con rimando allo standard ISO 639-2, come proposto dall'ICLTT [9], e la forma dell'etimo (vd. Fig. 3). Il rimando alla risorsa lessicografica, anch'esso elemento non strutturato, verrà codificato all'interno del valore <bibl> attraverso uno *script* di recupero automatico della risorsa etimologica di riferimento, più facilmente ricavabile rispetto alla sezione 'etimo testo libero' in quanto risponde a una minore variabilità di forme.

```
<dictScrap xml:id="TLIO.numerovoce.etim.">
   <etym>
    <seg type="desc"></seg>
    <cit type="etymon">
      <etym>
      <lang xml:lang="" sameAs="#ISO639-2"></lang>
      </etym>
      <form><orth></orth></form>
    </cit>
   </etym>
```

Figura 3. <etym>

La stessa natura non totalmente strutturata della sezione pone dei problemi anche per la codifica di 'certezza' all'interno dell'attributo `@cert` – per cui sono previsti quattro valori "high", "medium", "low" o "unknow", alla cui risoluzione si sta lavorando.
Una codifica differente è previsa per quei casi in cui l'etimo rimanda a un'altra voce, come per le voci di composti e derivati, per cui si prevederà un elemento di *cross-reference*. Tale codifica è oggi facilmente applicabile solo al *VSM*, in quanto risorsa totalmente digitale che prevede già dei collegamenti all'interno del DB fra voci di base e voci di derivati, ma si auspica l'applicazione anche alle altre risorse lessicografiche attraverso specifici *script* di collegamento. Per il *cross-reference* si utilizzerà l'indentazione dei seguenti elementi: `<xr><lbl><ref>` (vd. Fig. 4), come già stabilito da altri progetti [18, 19].

```
<xr type="related">
   <lbl>Cfr./Vd./Da</lbl>
   <ref type="etim" target="#TLIO.numerovoce"/>
</xr>
```

Figura 4. <xr><ref>

- `<usg>`

Una sezione specifica del *TLIO*, non prevista negli altri dizionari, è il punto 0.4 della voce, ovvero la 'distribuzione geolinguistica' nelle diverse aree italo-romanze a cui fanno riferimento i testi. La stessa è stata codificata attraverso l'elemento `<usg>` con indicazione di tipologia "geographic" e attributo `@xml:lang` per la varietà linguistica di riferimento, mentre il riferimento testuale verrà recuperato automaticamente (vd. Fig. 5), come in altre sezioni della voce, attraverso il rimando alla corrispondente sigla presente in GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini). Per tutti i dizionari compatibili e per le aree italo-romanze identificate dallo Standard ISO, saranno fornito l'aggancio allo Standard ISO 639-2 per le varietà linguistiche e allo standard ISO 3166 per le aree geografiche.

```
<dictScrap xml:id="TLIO.numerovoce.distrib.">
   <usg type="geographic" sameAs="#ISO3166">
      <lang xml:lang="" sameAs="#ISO639-2"></lang>
   </usg>
   <ref type="bibliography" target="#SIGLAGATTO">
      <bibl><author><title></title></author><date></date></bibl>
   </ref>
</dictScrap>
```

Figura 5. <usg>

- `<revisionDesc>`

Il modulo `<revisionDesc>`, previsto alla fine del `<teiHeader>`, verrà utilizzato per creare dei *backup* automatici delle voci nelle diverse revisioni, ovvero nel passaggio da uno stato di revisione a un altro, così da garantire il tracciamento delle versioni. Attraverso questo modulo, integrato nella piattaforma di gestione delle risorse lessicografiche, sarà possibile effettuare *query* specifiche per data di redazione o revisione delle voci. A differenza di quanto proposto da altri progetti [9] per cui è stato utilizzato il *tag* `<div>` con valore "revisionDesc", il nostro modello prevederà il modulo `<revisionDesc>`, data la creazione di *file* XML separati per ogni voce, così da dar conto dell'iter redazionale e dei possibili aggiornamenti sulla singola voce.

- <sense>

L'elenco delle definizioni e degli esempi sarà strutturato come in figura (vd. Fig. 6), attraverso specifici elementi per le marche d'uso, semantiche e metalessicografiche, il rimando alla fonte dell'esempio e l'aggancio automatico al punto 0.7 della voce, ovvero alla lista riepilogativa dei significati, attraverso un puntatore interno formato dagli elementi <xr> e

```
<dictScrap xml:id="TLIO.numerovoce.def">
   <sense n="1" xml:id="TLIO.numerovoce.def.numerodef" > <!-- def -->
   <usg type="domain"></usg> <!-- Marche d'uso -->
   <usg type="meaningType"></usg> <!-- Marche metalessicografiche o semantiche -->
   <gramGrp><gram type="pos"></gram><gram type="gender"></gram></gramGrp>
      <cit xml:id="TLIO.numerovoce.def.numerodef.es.numeroes" type="example" subtype="FC" n="1">
      <ref type="bibliography" target="#SIGLAGATTO">
      <bibl><author></author><title></title><date></date></bibl>
      <quote n=""></quote>
      </ref>
   </cit>
</sense>
</dictScrap>


<dictScrap xml:id="TLIO.numerovoce.listadef">
   <xr type="related"><ref type="sense" target="#TLIO.numerovoce.def.numerodef"></ref></xr>
</dictScrap>
```

Figura 6. <sense>

<ref>.

- <form>

Il *VSM* prevede uno specifico punto della voce (0.8) che testa la sopravvivenza del lessema nel dialetto siciliano attraverso lo spoglio di lessici seriori rispetto all'arco cronologico coperto da ARTESIA e, dunque, dal vocabolario [14]. Si tratta di un campo specifico per il *VSM* per cui è stata prevista una codifica attraverso l'elemento <form> e l'attributo @type con valore "lemma" e rimando alla risorsa lessicografica. La forma registrata sarà inserita nell'elemento <orth> con eventuale attributo @n e valore numerico per le diverse forme attestate nei dizionari seriori. I tre gruppi del punto 0.8 previsti dal *VSM* per distinguere le corrispondenze nei dizionari cinquecenteschi (dei lessicografi Lucio Cristoforo Scobar e Nicola Valla), nei dizionari siciliani dal '700 al '900 e nel *TLIO* verranno codificati all'interno dell'attributo @n con valore "A", "B", "C" per le tre sezioni (vd. Fig. 7).

```
<dictScrap xml:id="VSM.numerovoce.corrisp">
   <form xml:id="VSM.numerovoce.corrisp.SC" type="lemma" n="A">
      <orth></orth>
   </form>
   <form xml:id="VSM.numerovoce.corrisp.VS" type="lemma" n="B">
      <orth type="variant" n="1"></orth>
   </form>
   <form xml:id="VSM.numerovoce.corrisp.TLIO" type="lemma" n="C">
      <orth></orth>
   </form>
</dictScrap>
```

Figura 7. Lessicografia seriore

4. RISULTATI E PROSPETTIVE FUTURE

Il risultato della ricerca qui proposta è un modello di voce in XML/TEI che permette di importare ed esportare le voci sul *database* relazionale gestito dal sistema LexiCad/Pluto e di collegare le diverse imprese lessicografiche all'interno di un unico sistema di codifica. Inoltre, la codifica porterà alla conservazione e alla gestione delle diverse fasi redazionali della voce, attraverso *backup* automatici su un sistema diverso dal *database* principale, con il duplice vantaggio di non appesantire lo stesso e di poter archiviare le diverse versioni della voce durante le più importanti fasi di redazione e revisione. Si sta, inoltre, lavorando alla possibilità di fornire agli utenti un *file* XML/TEI scaricabile e utilizzabile per altri progetti, oltre che alla definizione di una serie di *query* messe a disposizione dell'utente e basate sugli elementi codificati. La conversione dell'XML delle voci del *TLIO* nello standard XML/TEI ha il vantaggio di fornire uno strumento che può essere riadattato e riutilizzato in progetti differenti, garantendo una piena interoperabilità, lo scambio reciproco di dati e la convergenza di metodi di trattamento condivisi.

In accordo con i progetti portati avanti da TEI Lex-0 e con la struttura onomasiologica proposta per il *VSM* sulla base dell'*Historical Thesaurus of English* (*HTE*)[11] [15: 180], la codifica qui proposta sarà la base per successive integrazioni in ambito ontologico, con la possibilità di sviluppare un'ontologia in grado di rispondere coerentemente alle categorie del mondo discusse in un *corpus* medievale e di legarsi al progetto *Ontolex-Lemon*[12], sondando la convergenza di metodi e prospettive fra dizionari onomasiologici e ontologie [10]. Nella costruzione di un'ontologia per le risorse lessicografiche dell'italiano antico, verrà presa in considerazione la struttura ricorsiva già avanzata [17] nel collegamento fra l'entrata lessicale, il senso primario, le definizioni e gli esempi, con la definizione di entità astratte superiori a cui collegare il lemma e le definizioni dello stesso.

BIBLIOGRAFIA

[1] Arcidiacono, Salvatore. *Lessicografia elettronica e italiano delle origini*. Palermo: Centro di studi filologici e linguistici siciliani, 2022.

[2] Arcidiacono, Salvatore. *«Percorsi di lessicografia computazionale per un Vocabolario del Siciliano Medievale (VSM)»*. Bollettino del Centro di studi filologici e linguistici siciliani 24 (2013): 87-108.

[3] Arcidiacono, Salvatore. *«Pluto. Piattaforma Lessicografica Unica delle Origini».* In Italiano antico, italiano plurale. Testi e lessico del Medioevo nel mondo digitale. Atti del Convegno internazionale in occasione delle 40.000 voci del TLIO (Firenze, 13-14 settembre 2018), (a cura di) Lino Leonardi e Paolo Squillacioti. Alessandria: Edizioni dell'Orso, 2019.

[4] Bański, Piotr, Jack Bowers, e Tomaz Erjavec. «TEI Lex0 guidelines for the encoding of dictionary information on written and spoken forms». In *Electronic Lexicography in the 21st Century. Proceedings of ELex 2017 Conference (Sep 2017, Leiden, Netherlands)*, (a cura di) Iztok Kosem, Carole Tiberius, Jelena Kallas, Simon Krek, Vít Baisa, e Miloš Jakubíček, 485-494. Brno: Lexical Computing CZO s.r.o., 2018.

[5] Beltrami, Pietro G. «Lessicografia e filologia in un dizionario storico dell'italiano antico». In *Storia della lingua e filologia. Atti del Convegno ASLI (Pisa-Firenze, 18-20 dicembre 2008)*, 235-248. Firenze: Cesati, 2010.

[6] Boccellari, Andrea. «Il sistema di redazione e pubblicazione web del TLIO». In *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori (Firenze, 26 ottobre 2010)*, 57-64. Supplementi al Bollettino dell'Opera del Vocabolario Italiano, 3. Edizioni dell'Orso, 2012.

[7] Boccellari, Andrea, e Domenico Iorio-Fili. «Il supporto dell'informatica al Vocabolario». In «Diverse voci fanno dolci note». *L'opera del Vocabolario Italiano per Pietro G. Beltrami*, (a cura di) Pär Larson, Paolo Squillacioti, e Giulio Vaccaro, 15-30. Alessandria: Edizioni dell'Orso, 2013.

[8] Bowers, Jack, Axel Herold, Toma Tasovac, e Laurent Romary. «TEI Lex-0 Etym: Toward Terse Recommendations for the Encoding of Etymological Information». *Journal of the Text Encoding Initiative*, fasc. Rolling Issue (2022). http://journals.openedition.org/jtei/4300.

[9] Budin, Gerhard, Stefan Majewski, e Karlheinz Mörth. «Creating Lexical Resources in TEI P5*». Journal of the Text Encoding Initiative* 3 (2012). http://journals.openedition.org/jtei/522.

[10] França, Patrícia Cunha. «Onomasiological dictionaries and ontologies». In *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6-10 July 2010)*, (a cura di) Anne Dykstra e Tanneke Schoonheim, 1291-98. Fryske Akademy, 2010.

[11] Ide, Nancy, Adam Kilgarriff, e Laurent Romary. «A Formal Model of Dictionary Structure and Content». In *Proceedings of the Ninth EURALEX International Congress. EURALEX 2000 (Stuttgart, Germany, August 8th–12th, 2000)*, 113-126. Stuttgart: Uversität Stuttgart, Institut für maschinelle Sprachverarbeitung, 2000.

[12] Kilgarriff, Adam. «Use of Computers in Lexicography». In *Encyclopedia of Language and Linguistics*, (a cura di) Keith Brown, 783-793. Amsterdam: Elsevier, 2006.

[13] Lemnitzer, Lothar, Laurent Romary, e Andreas Witt. «Representing human and machine dictionaries in Markup languages». In Dictionaries. *An International Encyclopedia of Lexicography*, (a cura di) Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, e Herbert E. Wiegand, Supplementary volume: Recent developments with special focus on computational lexicography: 1195-1209. Berlin-Boston: De Gruyter Mouton, 2014.

[14] Mosti, Rossella. *«Il Vocabolario del Siciliano Medievale (VSM): primi risultati, riflessioni e prospettive»*. Bollettino del Centro di studi filologici e linguistici siciliani 33 (2023): 155-191.

[15] Pagano, Mario, Tecla Chiarenza, e Salvatore Arcidiacono. «Lessico siciliano medievale e contemporaneo: note di lavoro». In *Dialetto, uno nessuno centomila,* (a cura di) Gianna Marcato, 173-184. Padova: Cleup, 2017.

[16] Romary, Laurent. «Standardization of the Formal Representation of Lexical Information for NLP». (a cura di) Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, e Herbert Ernst Wiegand, *Supplementary volume: Recent developments with special focus on computational lexicography*:1266-1274. Berlin-Boston, 2014.

[11] L'*HTE* è un ricchissimo tesauro che organizza l'intero lessico della lingua inglese su base semantica, a partire da tre nodi principali chiamati *megacategories*. Attraverso la struttura dell'*HTE* sono stati classificati tutti i significati dell'*Oxford English Dictionary* (*OED*). In rete: https://ht.ac.uk.

[12] https://www.w3.org/2019/09/lexicog/

[17] Romary, Laurent. «TEI and LMF crosswalks». *Journal for language technology and computational linguistics* 30, fasc. 1 (2015). http://www.jlcl.org.

[18] Romary, Laurent, e Werner Wegstein. «Consistent Modeling of Heterogeneous Lexical Structures». *Journal of the Text Encoding Initiative* 3 (2012). http://journals.openedition.org/jtei/540.

[19] Salgado, Ana, Rute Costa, Toma Tasovac, e Alberto Simões. «TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa». In *Electronic lexicography in the 21st century: Smart lexicography. Proceedings of eLex 2019 conference (Sintra, Portugal, 1-3 October 2019)*, 417-433. Brno: Lexical Computing, 2019.

[20] Zgusta, Ladislav. *A Manual of Lexicography*. Berlin-Boston: De Gruyter Mouton, 1971.

Analisi Computazionale, Intelligenza Artificiale e Linguistica

Analisi computazionale dei report di sostenibilità: la vaghezza come strategia di greenwashing

Erica Cutuli
Università di Catania, Italia – erica.cutuli@phd.unict.it

ABSTRACT

Oggi è possibile elaborare in maniera automatica enormi quantità di testo in linguaggio naturale, ma vi sono ancora difficoltà nel trattare casi di ambiguità e vaghezza linguistica. L'obiettivo del presente contributo è l'analisi dei fenomeni di vaghezza e imprecisione nei report di sostenibilità, con particolare attenzione al *greenwashing*.

Il corpus oggetto dell'analisi è composto da 225 report di sostenibilità in lingua italiana in pdf, pubblicati da 45 aziende, relativi agli anni tra 2017 il 2021. Questi sono stati analizzati ed elaborati con il tool Sketch Engine e un notebook in Python, partendo dalla ricerca mirata di parole chiave solitamente legate al *greenwashing*, e annotando un campione di concordanze. In particolare, per ciascun esempio estratto è stato stabilito se rispondesse ai criteri per essere considerato un'*asserzione ambientale* e se fosse vago, classificando ove possibile i casi di vaghezza in cinque categorie semantiche (*quantity*, *degree*, *time*, *category* e *softening stancetaking*).

Il contributo del lavoro è duplice, poiché da un lato l'analisi preliminare sul linguaggio utilizzato nei report di sostenibilità fornisce una migliore comprensione delle strategie linguistiche associate alle *asserzioni ambientali*, promuovendo una maggiore trasparenza e responsabilità da parte delle aziende; dall'altro pone le basi per l'identificazione automatica di quelle riconducibili al *greenwashing* con un dataset in lingua italiana per l'allenamento di modelli di Intelligenza Artificiale con un valore aggiunto in prospettiva del raggiungimento dell'uguaglianza linguistica digitale.

PAROLE CHIAVE

Natural Language Processing; sostenibilità; greenwashing; vaghezza; asserzioni ambientali.

1. CONTESTO E SCOPO DELLO STUDIO

Questo contributo si colloca tra le ricerche che utilizzano il *Natural Language Processing* (NLP) e l'analisi computazionale dei testi nell'ambito della sostenibilità. Esso si pone come obiettivo l'estrazione e l'analisi di asserzioni ambientali da report di sostenibilità aziendali e l'individuazione tra queste di quelle vaghe e riconducibili alla pratica del *greenwashing*. Per far ciò è stato compilato un elenco di termini facilmente correlabili al fenomeno, nell'ipotesi che l'osservazione delle concordanze di questi termini possa consentire l'estrazione di affermazioni ambientali di cui poter verificare il grado di vaghezza.

Prima di presentare la metodologia e i risultati, è necessario innanzitutto introdurre qualche definizione, partendo dai concetti di vaghezza linguistica e di greenwashing, e fornire sia una breve panoramica del contesto sia le motivazioni dietro alcune scelte metodologiche.

I concetti di vaghezza e di *greenwashing* condividono la caratteristica di essere multidisciplinari e per questo non consentono un'unica definizione ma molteplici a seconda di chi li studia e con che scopi. Per quanto riguarda la vaghezza, alcuni filosofi come Russell [10] sostengono sia una proprietà intrinseca del linguaggio, mentre alcuni linguisti ne sottolineano le finalità comunicative: "Il linguaggio vago è un fenomeno linguistico naturale, solitamente intenzionale e multifunzionale, che comporta imprecisioni e viene impiegato per determinate strategie comunicative." [11][1]. Queste strategie comunicative sono spesso sfruttate nell'ambito del marketing o, più in generale, della comunicazione aziendale, per rendere le pubblicità più persuasive [3] o costruire un'immagine positiva dell'azienda [5].

In questo contributo, l'analisi della vaghezza è ristretta all'ambito della sostenibilità, e in particolare al fenomeno del greenwashing, che può presentarsi sotto forma di: "[...] *greenwashing* a livello di prodotto/servizio, che utilizza argomentazioni testuali che si riferiscono esplicitamente o implicitamente ai benefici ecologici di un prodotto o servizio per creare un'affermazione ambientale fuorviante" [4][2] e, più nello specifico, come "affermazioni eccessivamente vaghe, ambigue, troppo ampie e/o prive di una chiara definizione" [4][3]. Questa pratica, genericamente descritta come il mostrarsi

[1] Questa e le successive citazioni dall'inglese sono state tradotte in italiano dall'autrice; per completezza si riportano in nota i testi originali. Testo in lingua originale: "vague language is a natural, usually purposeful and multi-functional linguistic phenomenon that involves imprecision and is employed for certain communicative strategies".

[2] "[…] product/service-level claim greenwashing, which uses textual arguments that explicitly or implicitly refer to the ecological benefits of a product or service to create a misleading environmental claim".

[3] "claims that are overly vague, ambiguous, too broad, and/or lacking a clear definition".

più sostenibili di quanto non si sia in realtà (per una rassegna dei diversi casi specifici si rimanda ad altri testi [1, 4]), rappresenta un ostacolo agli obiettivi per uno sviluppo sostenibile. Infatti, anche la Commissione Europea dichiara "[…] l'impegno a contrastare la problematica delle asserzioni ambientali false, garantendo agli acquirenti di ricevere informazioni attendibili, comparabili e verificabili, e così permettendo loro di prendere decisioni più sostenibili e ridurre il rischio di un marketing ambientale fuorviante (greenwashing)"[4]; dove per asserzione ambientale si intende "un messaggio o una dichiarazione […] che asserisce o induce a ritenere che un dato prodotto o professionista ha un impatto positivo o nullo sull'ambiente oppure è meno dannoso per l'ambiente rispetto ad altri prodotti o professionisti oppure ha migliorato il proprio impatto nel corso del tempo"[5].

Nonostante le direttive della Commissione Europea riguardino per lo più la comunicazione delle aziende verso i consumatori a scopi pubblicitari, si ritiene che gli stessi concetti possano essere applicati alla comunicazione verso gli *stakeholders* tramite i report di sostenibilità, dato che è possibile identificare strategie comunicative persuasive all'interno di tali report.

I report aziendali (nelle loro varie declinazioni come CSR, ESG, finanziari, non finanziari ecc.) sono oggetto di un filone di studi da diverse prospettive (linguistiche, comunicative, economiche, sociali…) e data l'enorme quantità di documenti prodotti, anche a seguito dell'obbligatorietà di tali report per determinate categorie di aziende, molti di questi progetti utilizzano il NLP per accedere ai dati testuali ed analizzarli. La rassegna proposta da Moodaley e Telukdarie [9] mostra infatti come l'applicazione di Intelligenza Artificiale (IA), NLP e *Machine Learning* ai report di sostenibilità sia già molto diffusa, mentre l'applicazione di queste tecniche nell'ambito del *greenwashing* e gli studi sul *greenwashing* in relazione ai report di sostenibilità sono dei campi di ricerca nascenti. Gli stessi autori hanno delineato un quadro concettuale per l'adattamento di *Large Language Models* (LLMs) al dominio specifico della sostenibilità [8]: Webersinke et al. propongono una versione del modello BERT allenato su testi inerenti all'ambito climatico [13]; Stammbach et al. propongono un dataset annotato e un modello allenato su questo dataset con l'obiettivo specifico di automatizzare l'individuazione di frasi legate all'ambiente [12].

Facendo riferimento a questi lavori, questo contributo si pone un secondo obiettivo, ovvero quello di costruire un dataset annotato in lingua italiana con lo scopo di essere utilizzato successivamente per l'allenamento di modelli di IA. Considerando che la quasi totalità delle ricerche in questo campo è effettuata sulla lingua inglese, riteniamo che lavorare sull'italiano abbia quindi un valore aggiunto ai fini del raggiungimento dell'uguaglianza linguistica digitale. Inoltre, è stato dimostrato che una comunicazione aziendale superficiale e non verificabile è correlata in modo significativo con un aumento delle emissioni, in quanto le aziende che più sfruttano questo tipo di comunicazione "danno priorità al mantenimento di una percezione pubblica positiva piuttosto che all'introduzione di cambiamenti significativi nelle loro pratiche aziendali." [2][6]. Dunque, l'individuazione di questi fenomeni linguistici potrebbe consentire un monitoraggio delle aziende più mirato e avere un impatto positivo verso il raggiungimento degli obiettivi di sostenibilità.

2. CORPUS E ANALISI

L'analisi è stata condotta su un corpus composto da 225 report di sostenibilità in lingua italiana, relativi a cinque anni consecutivi (2017-2021) e pubblicati da 45 aziende. Le aziende sono state identificate da un elenco pubblicato dalla Commissione Nazionale per le Società e la Borsa (CONSOB)[7], che comprendeva i soggetti che, nel periodo compreso tra il 1° gennaio e il 16 novembre 2022, avevano reso pubblici i loro bilanci non finanziari per l'anno fiscale iniziato il 1° gennaio 2021. Tra queste, ne sono state selezionate 45 in maniera casuale con la condizione che fossero disponibili i loro report per gli anni di interesse[8]. Considerando il fatto che ciascuna azienda può operare in settori diversi, si riporta in Tabella 1 un riepilogo dei settori rappresentati all'interno del corpus.

[4] Commissione europea. «Proposta di direttiva del Parlamento europeo e del Consiglio sull'attestazione e sulla comunicazione delle asserzioni ambientali esplicite (direttiva sulle asserzioni ambientali)» COM/2023/166 final, 22 marzo 2023. https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:52023PC0166.

[5] Vd. nota 5.

[6] "[…] prioritize maintaining a positive public perception over making meaningful changes to their business practices."

[7] Disponibile a questo URL: https://www.consob.it/web/area-pubblica/storico-elenco-dnf-al-16-novembre-2022

[8] Di seguito l'elenco completo delle aziende selezionate: Amplifon Spa, Arnoldo Mondadori Editore Spa, Atlantia Spa, Autogrill Spa, Avio Spa, Banca Generali Spa, Be Shaping The Future Spa, Brembo Spa, Brunello Cucinelli Spa, Buzzi Unicem Spa, Cairo Communication Spa, Caltagirone Spa, Cembre Spa, Datalogic Spa, Diasorin Spa, Dovalue Spa, El.En. Spa, Elica Spa, Enav Spa, Enel Spa, Fiera Milano Spa, Fincantieri Spa, Gruppo Mutuionline Spa, Immsi Spa, Iren Spa, Leonardo Spa, Lu-Ve Spa, Moncler Spa, Monrif Spa, Openjobmetis Spa, Pininfarina Spa, Pirelli & C. Spa, Prima Industrie Spa, Saes Getters Spa, Safilo Group Spa, Salvatore Ferragamo Spa, Saras Spa, Servizi Italia Spa, Sol Spa, Technogym Spa, Tesmec Spa, Tod's Spa, Unipol Gruppo Spa, Webuild Spa, Zignago Vetro Spa.

Settore	Numero di aziende che operano nel settore	Percentuale
Settore industriale	17	38%
Beni di consumo ciclici	13	29%
Finanza	9	20%
Servizi aziendali	9	20%
Servizi al consumatore	6	13%
Sanità	5	11%
Materiali non energetici	5	11%
Tecnologia	4	9%
Servizi pubblici	3	7%
Beni di consumo non ciclici	3	7%
Energia	2	4%

Tabella 1. Tabella riassuntiva dei settori in cui operano le aziende selezionate.

Il settore industriale, che include la produzione e la gestione di infrastrutture e tecnologie per il trasporto, emerge come il più numeroso. Segue il settore del commercio al dettaglio e della produzione, che spazia dall'abbigliamento e gli accessori alla produzione di componenti per veicoli e all'intrattenimento online. In contrasto, i settori legati alla produzione, distribuzione e vendita di energia, così come quelli relativi ai servizi per la salute, la cura personale e l'educazione, risultano meno presenti.

I report di sostenibilità sono stati scaricati in formato pdf dai siti ufficiali delle aziende e sono stati analizzati con il tool Sketch Engine [6], che permette facilmente l'estrazione del testo e il suo processamento (come ad esempio *tokenizzazione* e *Part Of Speech tagging*). L'estrazione di testo da pdf è un passaggio ancora problematico e comporta sempre un certo grado di imprecisione poiché, soprattutto in presenza di immagini e impaginazioni molto complesse, il riconoscimento dei diversi blocchi di testo può venir meno generando delle incoerenze. Il grado di accuratezza di Sketch Engine nell'estrazione del testo è stato ritenuto accettabile dall'autrice per gli scopi di questo studio, nonostante la divisione in *sentences* sia poco performante, probabilmente a causa della natura dei report che sono ricchi di titoli ed espedienti grafici utilizzati per separare le frasi a scapito della punteggiatura.

Una volta caricato e processato, il corpus contiene 14.279.823 *tokens* e 304.698 *sentences*. La distribuzione dei *tokens* per anno mostra un andamento di crescita costante passando da quasi due milioni (1.917.907) per i report del 2017 a quasi quattro milioni (3.858.273) per il 2021. Questo dimostra una chiara tendenza per le aziende a scrivere report di sostenibilità in media sempre più lunghi, con un contenuto testuale che arriva a raddoppiare nei cinque anni presi in esame. Per quanto riguarda la distribuzione dei *tokens* per azienda si passa da un massimo di 1.286.124 (Banca Generali Spa, 9% del corpus) ad un minimo di 60.291 (Gruppo Mutuionline Spa, 0,4% del corpus), cosa che rende alcune aziende più impattanti sull'analisi di altre.

Data l'assenza di dataset annotati o corpora composti esclusivamente da esempi di *greenwashing*, l'implementazione di un processo di estrazione automatica delle parole chiave non è stata possibile. Pertanto, si è scelto di fare riferimento alla metodologia utilizzata da Li [7] e alla documentazione della Commissione Europea[9], ed è stata stilata di conseguenza una lista di termini ritenuti facilmente associabili al fenomeno del *greenwashing*, cercando di mantenere una giusta rappresentazione della tematica ambientale e della vaghezza linguistica. L'ipotesi di partenza è che la ricerca di questi termini consenta di individuare *asserzioni ambientali* delle quali verificare l'eventuale vaghezza leggendone le concordanze. I parametri utilizzati per queste ricerche su Sketch Engine sono: "*simple*", poiché effettua la ricerca per lemma e non distingue tra minuscole e maiuscole; e "*sentences*", poiché l'obiettivo è estrarre le frasi contenenti le parole ricercate e non guardare il contesto generale. Sono stati utilizzati alcuni simboli speciali come nel caso di "eco--sostenibile" per includere nella ricerca le varie forme "ecosostenibile", "eco sostenibile" e "eco-sostenibile"; oppure nel caso di "emissioni zero|zero emissioni" per raggruppare nella stessa ricerca le due combinazioni.

I risultati ottenuti dalle ricerche delle concordanze sono riassunti nella Tabella 2. Questi sono stati scaricati in formato xlsx ed elaborati attraverso l'utilizzo di un notebook in *python*. Il *dataframe* che unisce tutti i risultati delle ricerche contava 35.515 righe, ridotte a 33.455 dopo l'eliminazione di quelle duplicate, dato che le parole ricercate possono ripetersi all'interno della stessa frase o in frasi limitrofe quando non divise correttamente dal processamento. È possibile che la

[9] Commissione europea. «Proposta di direttiva del Parlamento europeo e del Consiglio sull'attestazione e sulla comunicazione delle asserzioni ambientali esplicite (direttiva sulle asserzioni ambientali)» COM/2023/166 final, 22 marzo 2023. https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:52023PC0166

stessa azienda ripeta alcune frasi in report diversi e che una frase sia ripetuta in più righe perché contenente due o più termini ricercati, è stato dunque verificato il numero di frasi distinte pari a 27.888.
Dato il numero elevato di frasi estratte, si è scelto di scorporare dei campioni in maniera casuale per proseguire l'analisi di tipo qualitativo. Per la discussione dei diversi casi, si riportano di seguito tre esempi in cui sono state sottolineate le parti di interesse. Per ciascuna è stato annotato dall'autrice se si tratta di un'*asserzione ambientale* (*environmental claim*) e se possa essere considerata vaga (*vagueness*).

Ricerca	**Numero di occorrenze totali**	**Numero di frasi distinte**
sostenibile	7.831	6.951
riduzione	7.419	6.038
ridurre	4.149	3.426
co2	3.782	3.155
naturale	3.098	2.607
impatto ambientale	2.857	2.351
green	1.785	1.540
riciclare	1.595	1.341
biodiversità	1.020	905
verde	478	433
riciclato	357	327
ecologico	305	277
biologico	287	247
eco--sostenibile	118	106
emissioni zero|zero emissioni	112	106
riciclabile	99	79
minore impatto	82	72
eco--compatibile	70	59
biodegradabile	57	55
impatto zero	14	10
Totali	**35.515**	**27.888**[10]

Tabella 2. Tabella riassuntiva dei risultati di ricerca delle concordanze.

Esempio 1: "Un modello di business basato sulla circolarità implica la massima collaborazione tra tutti gli attori; per questo riteniamo fondamentale aprirci al confronto con i soggetti che condividono la nostra visione, coinvolgendo le filiere e promuovendo iniziative comuni per salvaguardare le risorse naturali e accrescere la competitività del Paese." - ENEL SPA, 2019

Esempio 2: "Quella di Montepacini è un'esperienza pluriennale di collaborazione pubblico-privato (TOD'S), finalizzata al pieno esercizio dei diritti delle persone disabili e fragili, che coinvolge volontari, associazioni e persone impegnate nelle tematiche relative a biodiversità , filiera corta, sostenibilità, solidarietà e valorizzazione delle produzioni locali "buone, pulite, giuste e per tutti"." - TOD'S SPA, 2020

Esempio 3: "Scatole e vassoi: anche per tale materiale di imballaggio, il Gruppo ha avviato una politica di acquisto orientata all'utilizzo di materiali riciclati, raggiungendo percentuali di assoluto interesse: infatti, nel 2020 oltre il 72% delle scatole e vassoi utilizzati sono stati prodotti con materiale riciclato." - ZIGNAGO VETRO SPA, 2020

Tutti e tre gli esempi possono essere considerati come *asserzioni ambientali* in quanto collegano l'azienda ad attività con impatto positivo sull'ambiente. Facendo riferimento a due analisi con obiettivi e metodologie simili al presente contributo, una su testi legislativi [7] e una su CSR reports [5], si è scelto di arricchire l'annotazione distinguendo dove possibile alcune

[10] Questa cifra si riferisce al numero totale di frasi distinte, rimuovendo tutte quelle duplicate per le diverse ragioni sopraindicate. Non corrisponde dunque alla somma delle cifre in colonna che rappresentano invece le frasi senza duplicati per ciascuna parola chiave ricercata.

caratteristiche linguistiche associate all'uso della vaghezza in relazione a cinque categorie semantiche: *degree, quantity*, *category, time* e *softening stancetaking*.
Degree: vengono inseriti in questo gruppo gli utilizzi di aggettivi e avverbi ambigui o di quei termini il cui significato dipende molto dal contesto e che necessiterebbero quindi di essere spiegati o esplicitati. Esempi come "massima" e "salvaguardare" (Esempio 1), "buone, pulite, giuste" (Esempio 2), "di assoluto interesse" (Esempio 3), sono tutti casi che non rispondono alle domande "come? / in che senso?". *Quantity*: rientrano in questa categoria le espressioni di approssimazione come "oltre il 72%" (Esempio 3) o "circa" oppure di quantità non specificate come "percentuali di assoluto interesse" (Esempio 3), e tutti i casi in cui non si trova risposta alle domande "quanto? / di quanto?" come nell'espressione "che coinvolge volontari" (Esempio 2) o nelle molte occorrenze relative alle ricerche di "ridurre/riduzione/minore impatto" che non includono specificazioni di quantità. *Category*: espressioni come "le risorse naturali" (Esempio 1), "persone disabili e fragili" (Esempio 2) e "materiali/e riciclati/o" (Esempio 3) sono associati ad un senso di vaghezza di questo tipo in quanto sembrano riferirsi ad una categoria di entità piuttosto che un oggetto specifico. *Time*: fanno parte di questa categoria espressioni come "pluriennale" (Esempio 2), "nel tempo" e "da anni". *Softening stancetaking*: nell'Esempio 1 possiamo notare che l'espressione "riteniamo fondamentale aprirci al confronto" è un modo più negoziabile per dire che l'azienda "si confronta"; sottolinea una presa di posizione ma senza utilizzare un'affermazione chiara, lasciando quindi il dubbio sull'azione effettivamente svolta dall'azienda.
La Tabella 3 riporta i risultati dell'annotazione di un *sample* di 100 frasi, proveniente da 65 documenti diversi, prodotti da 33 aziende diverse, leggermente sbilanciato in quanto 4 aziende costituiscono da sole un terzo del campione.

		vagueness	**quantity**	**degree**	**time**	**category**	**softeneng stance-taking**
not environmental claim	20	9	7	3	1	1	1
environmental claim	80	69	32	36	6	15	4
Totale complessivo	**100**	78	39	39	7	16	5

Tabella 3. Tabella riassuntiva dei risultati dell'annotazione di un campione di frasi.

L'ipotesi che i termini ricercati riuscissero ad individuare molti casi di asserzioni ambientali vaghe è stata confermata. La distribuzione sproporzionata tra le varie categorie semantiche è dovuta molto probabilmente alla scelta delle parole chiave in fase di selezione delle frasi.

3. SVILUPPI FUTURI

Dato che la condivisione dei dati è fondamentale per facilitare la riproducibilità dei risultati e per stimolare ulteriori ricerche, il dataset annotato utilizzato in questo studio sarà reso disponibile attraverso una repository pubblica su GitHub. Tra le prospettive future ci si propone prima di tutto di ampliare la lista dei termini ricercati. L'aggiunta di pattern linguistici e termini di natura vaga, come quelli che hanno determinato il grado e la tipologia di vaghezza negli esempi citati (ad esempio "di assoluto interesse", "circa", "pluriennale"), permetterà di ottenere una panoramica più estesa sulle strategie di vaghezza impiegate all'interno dei report e consentirà inoltre di identificare eventuali termini legati all'ambiente che erano stati precedentemente esclusi dalle ricerche.
Il processamento del testo all'interno dei file pdf può essere migliorato per permettere un'analisi più accurata: se da un lato i report sono testi ricchi di titoli, didascalie e espedienti grafici utilizzati per separare blocchi di testo, il processamento tramite Sketch Engine [6] pecca nella suddivisione delle frasi in mancanza di punteggiatura e talvolta nella gestione separata di colonne di testo; inoltre, non permette di distinguere frasi che fanno parte di un paragrafo più ampio da quelle utilizzate come titoli o didascalie, informazione che sarebbe utile nell'ottica di una ulteriore verifica dei casi di vaghezza. Il dataset taggato è molto piccolo ed il lavoro è stato effettuato da un solo annotatore, ci si propone dunque di aumentare la grandezza e l'accuratezza del dataset coinvolgendo più annotatori.

4. CONCLUSIONI

Dall'analisi effettuata emerge come le parole chiave selezionate per la ricerca di esempi di *asserzioni ambientali* vaghe siano utilizzate frequentemente all'interno del corpus nelle modalità ipotizzate. In particolare, i casi più frequenti sono quelli in cui le aziende utilizzano termini vaghi come "ecologico" o "sostenibile" senza fornire ulteriori informazioni all'interno della stessa frase che ne esplicitino il significato, con la conseguenza che le posizioni, le attività e i progressi delle aziende restano difficili da identificare con chiarezza. La seconda strategia più utilizzata è legata all'utilizzo di verbi come "ridurre, aumentare" o aggettivi come "minore, maggiore, ridotto" senza ulteriori specificazioni sulle quantità, oppure, quando una quantità viene specificata attraverso approssimazioni come "circa, fino a, oltre". Così facendo, le

aziende possono parlare dei propri (ipotetici) impatti positivi sull'ambiente, senza però dare una reale misura di questi, rendendo le informazioni difficilmente comparabili e verificabili.
In conclusione, si ritiene che studi come questo possano aiutare ad individuare i punti critici della comunicazione aziendale in materia di sostenibilità con l'obiettivo di favorire dichiarazioni più trasparenti e in linea con le direttive europee. Inoltre, la costruzione di dataset accuratamente annotati rappresenta un contributo importante per l'identificazione automatica del *greenwashing*, ancor di più se in lingua italiana in prospettiva del raggiungimento dell'uguaglianza linguistica digitale.

BIBLIOGRAFIA

[1] Bernini, Francesca, e Fabio La Rosa. «Research in the greenwashing field: concepts, theories, and potential impacts on economic and social value» *Journal of Management and Governance*, 2023.

[2] Bingler, Julia, Mathias Kraus, Markus Leippold, e Nicolas Webersinke. «How Cheap Talk in Climate Disclosures relates to Climate Initiatives, Corporate Emissions, and Reputation Risk» *Swiss Finance Institute Research Paper*, n. 22 (2022).

[3] Chen, Minghui. «Analysis of Vagueness in English Advertisement from the perspective of Adaption Theory» *International Journal of Social Science and Economic Research 3*, n. 3 (2018): 1068-1086.

[4] de Freitas Netto, Sebastião Vieira, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, e Gleibson Robert da Luz Soares. «Concepts and forms of greenwashing: a systematic review» *Environmental Sciences Europe* 32, n. 19 (2020).

[5] Jin, Bixi. «A corpus-assisted study of vague language in corporate responsibility reports of the cosmetics industry» *Ibérica* 43 (2022): 77-102.

[6] Kilgarriff, Adam and Baisa, Vít, et al. «The Sketch Engine: Ten Years On» *Lexicography 2197-4306* 1 (2014): 7–36.

[7] Li, Shuangling. «A corpus-based study of vague language in legislative texts: Strategic use of vague terms» *English for Specific Purposes* 45 (2017): 98-109.

[8] Moodaley, Wayne, e Arnesh Telukdarie. «A Conceptual Framework for Subdomain Specific Pre-Training of Large Language Models for Green Claim Detection» *European Journal of Sustainable Development* 12 (2023): 319.

[9] Moodaley, Wayne, e Arnesh Telukdarie. «Greenwashing, Sustainability Reporting, and Artificial Intelligence: A Systematic Literature Review» *Sustainability* 15 (2023): 1481.

[10] Russell, Bertrand. «Vagueness»: In *Vagueness: A reader* (a cura di) Rosanna Kenney, Peter Smith, 61-68. Cambridge: MIT Press, 1997.

[11] Ruzaitė, Jurate. *Vague language in educational settings: Quantifiers and approximators in British and American English.* Peter Lang, 2007.

[12] Stammbach, Dominik, Nicolas Webersinke, Julia Bingler, Mathias Kraus, e Markus Leippold. «Environmental Claim Detection» September 2022. Available at SSRN: https://ssrn.com/abstract=4207369.

[13] Webersinke, Nicolas, Mathias Kraus, Julia Bingler, e Markus Leippold. «CLIMATEBERT: A Pretrained Language Model for Climate-Related Text» *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*. 2022.

Analisi stilometrica applicata alle capacità emulative di GPT-4

Marco De Cristofaro[1], Mariangela Giglio[2]
[1]Université de Mons, C2W, European Postdoctoral programme Marie Curie Cofund Action, Belgio - marco.decristofaro@umons.ac.be
[2]Università di Bologna, Italia - mariangela.giglio2@unibo.it

ABSTRACT

La presente ricerca si propone di esplorare la capacità del Large Language Model GPT-4 di emulare lo stile di due riviste culturali italiane attive negli anni '60, «Quaderni Piacentini» e «Quindici». Utilizzando un corpus derivato dai primi numeri delle riviste, trascritti attraverso OCR e poi accuratamente rivisti, il lavoro vuole valutare se GPT-4 sia in grado di eludere l'analisi stilometrica, scrivendo un testo che rispecchi la strategia editoriale di una determinata rivista. Dopo aver utilizzato GPT-4 per generare testi che riflettessero lo stile delle riviste grazie a specifici prompt, è stata condotta un'analisi stilometrica per confrontare i testi artificialmente generati con il corpus originale. I risultati evidenziano le sfide e le potenzialità di GPT-4 nella creazione di scritti con stili complessi. Il confronto con metodologie stilometriche tradizionali ha permesso di individuare i punti in cui le due riviste si distanziano e, di conseguenza, i nodi cruciali su cui il modello si concentra per la differenziazione stilistica e tematica. La ricerca intende aprire nuove prospettive sull'utilizzo della stilometria per l'analisi computazionale di testi relativi a specifici contesti culturali; se fino ad ora, infatti, la comunità scientifica si è concentrata sulla capacità degli LLM di riprodurre fedelmente stili di diversi autori o sull'indagine di piattaforme di scrittura, applicare queste metodologie all'ambito ristretto delle riviste consentirebbe più ampie considerazioni sulle strategie editoriali, sulle tendenze di lettura e sui processi di circolazione delle idee.

PAROLE CHIAVE

Stylometry; Authorship attribution; Large Language Models; GPT-4; ChatGPT.

1. INTRODUZIONE

Le riviste letterarie, e più genericamente quelle culturali, hanno spesso rappresentato per autori e intellettuali un terreno di confronto e soprattutto un mezzo attraverso cui prendere posizione nel dibattito pubblico. Nel secondo Novecento esse si sono sempre più configurate come «istituzioni» [9] capaci di offrire non solo un canale di accesso al campo di produzione culturale [3], ma anche uno spazio di partecipazione in costante movimento. In un simile orizzonte, negli anni Sessanta in Italia un ruolo di primo piano è stato ricoperto da due riviste, capaci di coniugare la dimensione culturale con precise ambizioni politiche: «Quaderni Piacentini», fondata da Piergiorgio Bellocchio e Grazia Cherchi nel 1962[1], e «Quindici», fondata a Roma nel 1967 dal Gruppo 6[2]. Se diverse indagini sono state condotte sulla loro attività, sull'organizzazione redazionale, sulla loro posizione politico-culturale, ad oggi non risultano analisi computazionali dei testi pubblicati sulle loro pagine. Uno studio ad ampio raggio sullo stile degli articoli, sulle tematiche affrontate e sulle posizioni assunte dai diversi collaboratori fornirebbe non solo un quadro più completo sui punti di vista dei due giornali, ma anche una più profonda comprensione delle rispettive strategie editoriali, che avrebbe implicazioni sulle ricerche in merito alle modalità di lettura del pubblico. Il ricorso a metodologie stilometriche e a Large Language Models (LLM), come GPT-4, potrebbe così offrire un importante contributo, da una parte alle indagini sulla storia culturale italiana degli anni Sessanta, dall'altra alle ricerche più aggiornate sulla storia della lettura e della stampa.

Il presente lavoro si inserisce in questo contesto, con l'obiettivo di esplorare e valutare le capacità della stilometria nell'ambito dell'individuazione di specificità stilistiche e tematiche di ciascuna rivista, e nel contempo di riflettere sulle performance di GPT-4 applicate, più che sullo stile di singoli autori, su un intero corpo editoriale. La ricerca si propone di andare oltre i tradizionali ambiti di applicazione della stilometria, concentrando l'attenzione sulla capacità degli LLM di replicare la somma stilistica di queste pubblicazioni.

Il saggio prende spunto dallo studio di Simone Rebora [16], che ha esaminato la capacità dei modelli linguistici di grande scala, come GPT-3, di "ingannare" le tecniche stilometriche nell'attribuzione autoriale. Tuttavia, mentre Rebora si è concentrato sulla capacità emulativa di singoli autori, questa ricerca si distingue per l'analisi di un fenomeno più specifico,

[1] Sui «Quaderni Piacentini» si vedano almeno le due antologie: *"Quaderni piacentini" 1962-1968* [2] e *Prima e dopo il '68* [8]. Per una ricostruzione storica delle vicende della rivista si rimanda a [14, 15].
[2] Per le coordinate storiche di «Quindici» si rimanda all'antologia *Quindici: una rivista e il Sessantotto* [1].

tentando di esaminare la profondità e l'efficacia con cui GPT-4 può assimilare e riflettere lo stile editoriale di riviste culturali.

2. CREAZIONE DEL CORPUS DI RIFERIMENTO

Una prima complessità all'interno del nostro studio riguarda la scelta del corpus di riferimento. È evidente, infatti, come, in virtù del loro ruolo chiave nella diffusione delle idee e nel dibattito pubblico, le riviste italiane del Novecento siano tanto numerose quanto variegate. Dal momento che spesso esse condividono scopi politico-culturali, forme espressive (racconti, reportage, articoli di critica ecc.) e approcci (commistione tra testi e immagini, utilizzo di traduzioni, coinvolgimento dei lettori) ci siamo chiesti se fosse possibile delimitare linee di demarcazione tra le diverse identità delle singole riviste. Da questa prima necessaria verifica discende la seconda criticità del nostro studio: trattandosi di testi di autori diversi, è possibile identificare elementi ricorrenti tra i vari contenuti di una sola rivista tali da determinarne la strategia editoriale? L'individuazione dei due giornali di riferimento dovrebbe rispondere, dunque, a questi specifici criteri di indagine. La scelta è ricaduta su «Quaderni Piacentini» e «Quindici» per diverse ragioni. In primo luogo, entrambe condividono, come emerge fin dalle dichiarazioni programmatiche dei rispettivi numeri iniziali, l'ambizione a considerarsi ad un tempo una rivista culturale ma anche uno spazio di intervento politico. Il secondo motivo è la prossimità temporale: è vero, infatti, che il primo numero di «Quaderni Piacentini» esce nel 1962 e quello di «Quindici» nel 1967, ma entrambi affrontano tematiche riconosciute come rappresentative del contesto sociale degli anni Sessanta prima del 1968. Inoltre, la somiglianza delle tematiche avrebbe garantito una maggiore centralità dell'analisi prettamente stilistica quale tratto dirimente della polarizzazione tra le due riviste: l'indagine stilometrica, in altre parole, sarebbe stata maggiormente valida, perché basata su aspetti stilistici e non influenzata da un'eccessiva distanza degli argomenti trattati. Infine, abbiamo deciso di concentrarci sui numeri iniziali delle riviste perché essi contengono dichiarazioni programmatiche, idealmente finalizzate a offrire al lettore una linea editoriale.

2.1 OCR e accuratezza

Una volta identificati i testi di riferimento si è provveduto alla trascrizione degli stessi: per quanto riguarda «Quaderni Piacentini» ci si è basati sull'antologia a cura di Luca Baranelli e Grazia Cherchi; per «Quindici» ci siamo affidati al volume curato da Nanni Balestrini. Sono stati dunque trascritti nell'interezza i primi tre numeri di ciascuna rivista per un totale di circa 25mila parole per «Quindici» e 22mila circa per «Quaderni Piacentini»[3]. Per garantire livelli di accuratezza quanto più possibile elevati è stato scelto di combinare tecniche OCR e forme di controllo manuale. Per la fase di riconoscimento OCR sono stati utilizzati congiuntamente due strumenti diversi: la piattaforma *Transkribus* e lo *Snipping Tool* sviluppato per Windows 11. Il risultato della trascrizione OCR *raw* (senza revisione umana) è stato di un tasso di accuratezza di circa il 99% (un errore ogni 101 caratteri). A seguito della correzione il corpus di riferimento non presenta errori immediatamente rilevabili.

Come ricordato da Holley [11] l'accuratezza media dell'OCR senza correzione può variare fra il 71% e il 98,02% Al livello 71% si troverebbero 145 caratteri erronei in un paragrafo medio di 500 caratteri. Visto da un'altra prospettiva, ciò significa che il 29% del paragrafo sarebbe scorretto: ne risultano difficoltà per la corretta lettura e interrogazione. La soglia minima di accuratezza consigliata dalla DFG (Deutsche Forschungsgemeinschaft) per le applicazioni di ricerca è di almeno il 99,95%. La trascrizione a quattro mani garantisce un'accuratezza del 99,997% [10]; in questo caso si è optato per una combinazione di testo OCR e di revisione manuale che permette di raggiungere percentuali di accuratezza paragonabili a quelle della trascrizione a quattro mani in quanto rappresenta un processo di doppia correzione del risultato della scansione OCR [12].

3. LLM E COMPETENZA EMULATIVA

ChatGPT è un chatbot rilasciato da OpenAI il 30 novembre 2022 che usa tecniche di NLP per generare dialoghi con l'utente su un'ampia varietà di argomenti. Se da un lato sistemi di chatbot come ChatGPT possono essere validi alleati per il ricercatore, d'altro canto le problematiche legate a un loro uso massiccio emergono sempre con maggior chiarezza. ChatGPT (basato sulla tecnologia *Generative Pre-trained Transformer*) è in grado di generare testi simili a quelli umani in un'ampia varietà di stili e di lingue, con una fluenza tale che la maggior parte delle persone non riesce a distinguere tra testi generati da AI e quelli scritti da esseri umani [13, 17]. Si è lungamente discusso della difficoltà di identificare con certezza testi scritti da AI, tanto che OpenAI il 31 gennaio 2023 ha rilasciato un classificatore di testi AI. È la stessa OpenAI, tuttavia, ad ammettere i limiti dell'accuratezza di rilevamento di questo classificatore, che presenta oltretutto

[3] Trattandosi di un lavoro che rientra in una ricerca più ampia, il corpus revisionato sarà reso accessibile, il prima possibile, all'interno del seguente repository github: https://github.com/MarcoDeCristofaro/RivisteAnni60.git.

prestazioni più basse per le lingue diverse dall'inglese. Sembra particolarmente interessante, dunque, valutare l'effettiva capacità emulativa di GPT in relazione non tanto al singolo autore (come già esplorato da Rebora [16]) ma rispetto alla somma stilometrica di un'intera rivista, seppure con uno stile peculiare e ampiamente riconoscibile. Si è scelto di utilizzare, a questo fine, il modello GPT-4 per la sua aumentata capacità di calcolo unitamente alla possibilità di allegare blocchi testuali direttamente nell'interfaccia di chat.

4. DEFINIZIONE DEL TEST SET TRAMITE GPT-4

La definizione del test set ha rappresentato una fase cruciale per valutare l'abilità di GPT-4 nell'attività emulativa. Il corpus delle riviste è stato accuratamente suddiviso in blocchi di testo, ciascuno contenente un minimo di 5.000 parole. La soglia è stata stabilita in base a criteri stilometrici consolidati [6], che identificano questa lunghezza come necessaria per garantire analisi significative e affidabili. Ogni blocco di testo è stato presentato a GPT-4 tramite l'interfaccia di chat. L'intento era di sfruttare le avanzate capacità di elaborazione del linguaggio del modello per generare articoli che non solo rispecchiassero lo stile, ma anche le tematiche intrinseche dei testi originali. Questo processo si è tuttavia rivelato non privo di sfide. Le limitazioni intrinseche di GPT-4, in particolare per quanto riguarda il limite di lunghezza dei testi generabili, hanno imposto restrizioni significative; è inoltre emersa una notevole variabilità nei risultati ottenuti utilizzando lo stesso prompt in sessioni di chat differenti. Per quanto si sia cercato di non suggestionare eccessivamente l'elaborazione da parte di GPT, è stato necessario modificare leggermente i parametri per impedire sia l'esatto ripetersi dei risultati sia la creazione di testi poco coerenti. Seguendo le raccomandazioni della stessa *API section* di OpenAI[4] si è operato principalmente e in maniera alternata sulla temperatura[5] e top p[6]. Per il primo test-set di testi si è settata una temperatura di 0.8. Questo parametro, tuttavia, creava dei testi eccessivamente simili tra loro, difficilmente utilizzabili per le analisi stilometriche. Dopo una serie di prove si è scelto dunque di mantenere la temperatura consigliata di 1.0, la stessa utilizzata nell'interfaccia online di GPT 4. Per quanto riguarda il top p, invece, si è scelto di rimanere nei valori consigliati tra 0.7 e 0.9. Valori più bassi (ad esempio, intorno a 0.5) hanno prodotto risposte apparentemente più interessanti ma che, secondo il nostro giudizio, risultavano eccessivamente vicine alla mera riformulazione del testo proposto in partenza. Sarebbe stato utile poter giudicare come il modificarsi dei parametri potesse influenzare le analisi stilometriche ma, a causa della corposità dei testi richiesti, questo non è stato possibile. In generale si è tentato comunque di minimizzare l'influenza dell'aspetto umano nell'analisi stilometrica, garantendo così una maggiore oggettività nei risultati ottenuti. Lo stesso criterio è stato mantenuto nella stesura del prompt che si è cercato di rendere il più generico e imparziale possibile.

Il prompt scelto per la generazione di articoli è stato il seguente: «*I submitted to you a word file with a few italian texts. Write a text in italian of 1000 words in the style of the texts I submitted. Try to use the same stylistic features, to adress similar topics, take the same political and cultural positions taken in the articles you analized. Try to emulate the articles and mirror the style, stylistic feature as much as possible. language of the results "Italian" lenght of the results: "1000 words" minimum*».

Si è già lungamente discussa l'incapacità di GPT di contare (non soltanto i token), ma si è scelto comunque di mantenere un'indicazione di massima all'interno del prompt. La generazione testuale è stata ripetuta fino a raggiungere i testi della lunghezza desiderata. I testi così ottenuti sono stati poi uniti in un singolo file di circa 5000 parole l'uno.

La prima considerazione riguarda la scelta del prompt: in questo task specifico si è optato per un approccio *Zero-Shot Prompting*, in quanto a seguito di approcci *Few-Shots* è emersa una tendenza a contaminare i risultati della generazione con i prompt utilizzati. Si è già accennato, inoltre, alla difficoltà di ottenere output consistenti con lo stesso prompt. La risposta a una simile criticità è in parte imputabile al sistema di funzionamento di GPT, che sfrutta un approccio di tipo probabilistico per la generazione delle risposte. Si è deciso comunque di mantenere lo stesso prompt per la generazione di tutti i testi al fine di ottenere uno standard di riproducibilità quanto più possibile elevato.

Nonostante le difficoltà evidenziate il processo ha portato alla redazione di un totale di otto articoli, quattro per ciascuna delle due riviste. Questi testi formano il test set che rappresenta il nucleo centrale dell'analisi stilometrica condotta in questo studio.

[4] https://platform.openai.com/docs/api-reference/chat/create

[5] La temperatura è un parametro utilizzato per regolare la distribuzione delle probabilità dei possibili token generati da un modello di linguaggio, modificando i punteggi associati alle parole, chiamati logit. Più alta è la temperatura, più uniforme diventa la distribuzione delle probabilità delle parole, favorendo scelte meno probabili e risposte più creative. Al contrario, una temperatura più bassa accentua le differenze tra i logit, favorendo le scelte più probabili e producendo risposte più coerenti ma potenzialmente meno creative.

[6] Il parametro "top-p", conosciuto come "nucleus sampling", regola la distribuzione delle probabilità durante la generazione del testo, limitando le scelte del modello ai token con probabilità cumulativa più alta. Riducendo il valore di "top-p", si concentra sui token più probabili, garantendo risposte coerenti ma potenzialmente meno creative. Aumentando "top-p", si amplia la varietà di token considerati, ma a discapito della coerenza.

5. ANALISI STILOMETRICA

Una prima indagine è stata dunque effettuata su un insieme di otto testi, suddivisi in quattro testi per le due riviste. Occorre precisare che i testi non corrispondono a uno specifico articolo ma rappresentano una somma di articoli differenti. Considerando, infatti, che il numero minimo di parole per un'analisi stilometrica affidabile è di circa 5000, abbiamo ritenuto necessario raccogliere gli articoli dei primi due numeri di ciascuna rivista in quattro differenti insiemi. In questo modo abbiamo ottenuto un testo sufficientemente lungo per l'analisi stilometrica e allo stesso tempo capace di rispecchiare una visione editoriale specifica e non frammentata della rivista.

Per la prima fase ci siamo affidati all'applicazione di tecniche stilometriche attraverso il linguaggio di programmazione R, sulla base della metodologia sviluppata da Maciej Eder e Jan Rybicki [7]. La metrica di riferimento iniziale è stata "Delta" [5]. Attraverso l'utilizzo di una serie di *features* differenti, siamo stati in grado di individuare possibili traiettorie distintive tra le due riviste. Per la visualizzazione dei primi risultati ci siamo affidati alla creazione di dendogrammi [18], che permettono di raccogliere i testi in gruppi separati sulla base del grado di somiglianza: il raggruppamento consente di osservare anche i diversi livelli a cui i vari contenuti sembrano avvicinarsi.

Dopo questa prima fase, abbiamo portato avanti un confronto tra le riviste. Applicando la Zeta Analysis [5], abbiamo indagato la probabilità con cui determinate scelte lessicali possano rappresentare la visione delle sedi di pubblicazione.

5.1. Un confronto stilistico-tematico: «Quaderni Piacentini» e «Quindici»

Una prima analisi è stata necessaria a rilevare il numero di *most frequent words* (MFW), capace di offrire i risultati più stabili. È stata impostata la *feature* di incremento progressivo "mfw.incr=50", individuando nelle 150mfw il valore di maggiore stabilità.

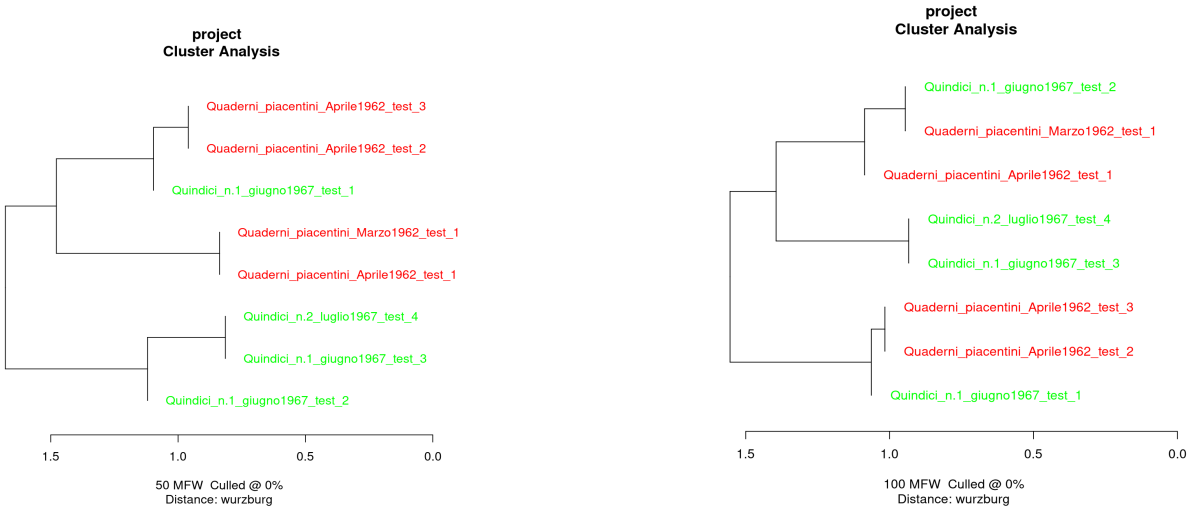


Figura 1. Cluster Analysis sul training set 50-100MFW

I risultati (vd. Fig. 1- Fig. 2) mostrano come le riviste si siano disposte sostanzialmente in due rami differenti evidenziando approcci distinti. Due casi significativi di contaminazione a livello stilometrico sono rappresentati dai testi estratti dalla seconda parte del secondo numero di «Quaderni Piacentini», pubblicato nell'aprile del 1962, e dalla seconda parte del primo numero di «Quindici», pubblicato nel giugno 1967. Procedendo con un'ulteriore analisi di comparazione con il metodo Zeta (vd. Fig. 3), che mette in opposizione a livello lessicale e in modo specifico i testi relativi alle due riviste, i risultati evidenziano una ben più chiara distanza rispetto ai termini maggiormente utilizzati. Nella Zeta Analysis le scelte lessicali individuano approcci ben più marcati, tracciando con maggior nettezza la posizione di ciascuna rivista. Di particolare rilievo in questo caso è la distanza tra i campi semantici preferiti dall'una e dall'altra rivista: mentre «Quaderni Piacentini» manifesta un maggior ricorso a termini di ambito politico come "lotta", "comunisti", "democrazia", "compagni", "socialista", "fascista", «Quindici» riflette gli interessi culturali dei suoi fondatori prediligendo una terminologia tipica della critica letteraria con parole quali "ipotesi", "lettore", "avanguardia".

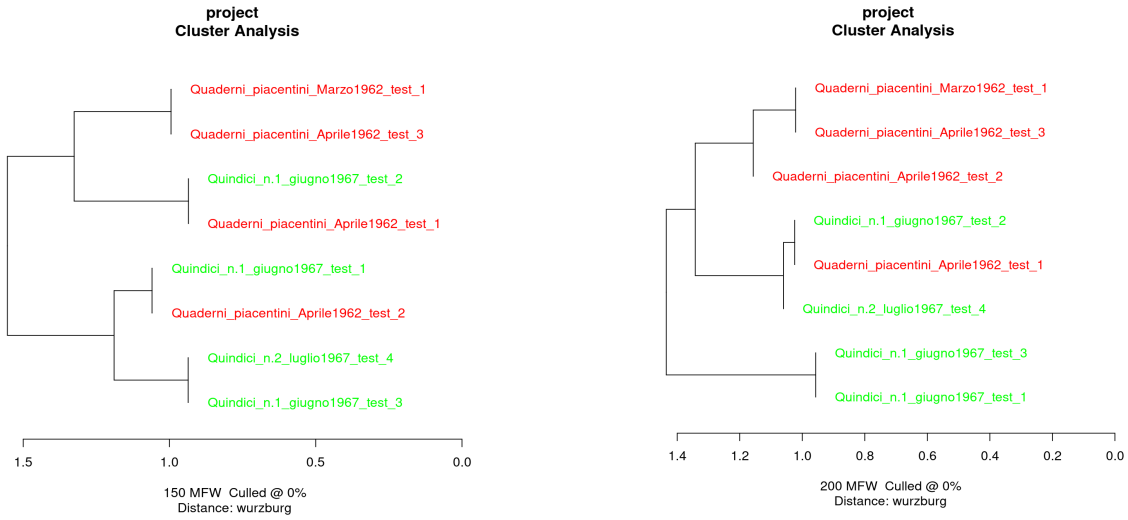


Figura 2. Cluster Analysis sul training set 150-200MFW

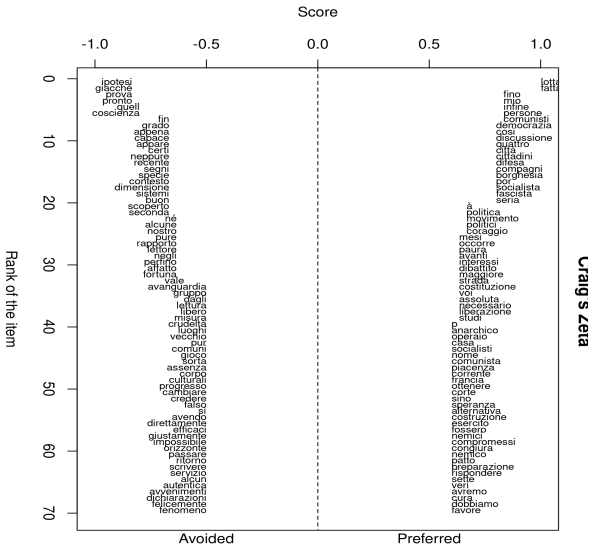


Figura 3. Zeta Analysis tra "Quaderni Piacentini" e "Quindici"

Se, dunque, un primo approccio complessivo sembra evidenziare una maggiore prossimità tra le due visioni politico-culturali, l'indagine lessicale esprime una più evidente contrapposizione, che sarà fondamentale tener presente al momento del confronto tra i testi originali e quelli prodotti da Large Language Models.

5.2. Stilometria su testi *AI generated*

Una volta appurata la specificità stilistica delle singole riviste, è stato possibile inserire nell'analisi anche i testi generati da GPT-4. Si sono dunque creati due blocchi testuali contenenti l'interezza degli articoli scritti da GPT per ogni rivista, in modo da poter efficacemente valutare il lavoro emulativo.

I risultati mostrano chiaramente la riconoscibilità della cifra stilistica dei testi prodotti con GPT-4. Alla bipartizione dell'albero già evidenziata in Fig. 1 si affianca un terzo ramo che pone in evidente separazione i testi *AI generated,* mantenendo visibile la contaminazione di cui sopra (vd. Fig. 4)

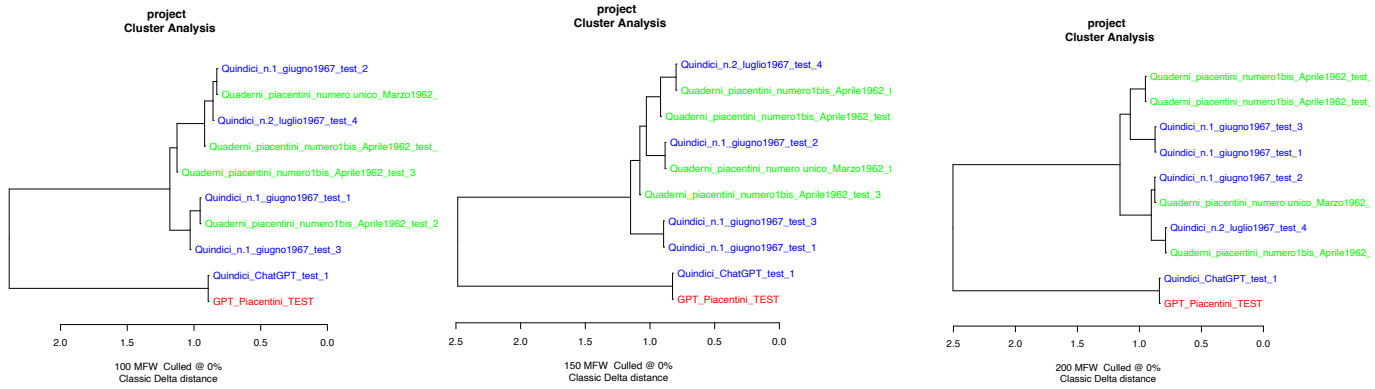


Figura 4. Cluster Analysis riviste e testi GPT-4

Aumentando il numero delle *most frequent words* si nota in modo ancora più evidente la corretta attribuzione. Non diversi sono i risultati con l'applicazione della Zeta Analysis (vd. Fig. 5). Quest'ultima mostra una separazione a livello lessicale tra i due insiemi di testi tanto più significativa se ci soffermiamo sui termini maggiormente utilizzati da ChatGPT in relazione alla scrittura delle riviste. Tra le parole che determinano il divario a cui ricorre l'LLM troviamo: società, ruolo, sfide, cambiamento, contesto, arte, futuro, culturale, sociale. Si tratta di termini portatori di significato e dal grande peso

semantico. Se essi risultano tanto utilizzati da GPT in relazione alle riviste, si può ipotizzare che il modello, nel tentativo di imitare lo stile dei giornali culturali, si affidi maggiormente ai nuclei tematici dei testi più che alle function words. L'aspetto tematico di conseguenza prevale su quello stilistico nell'addestramento del modello ma è, allo stesso tempo, il principale artefice del suo fallimento imitativo: lo sforzo di concentrarsi, infatti, sulle parole significative non è premiato, anzi, appare come l'elemento che causa la distanza tra i due approcci alla descrizione della realtà circostante.

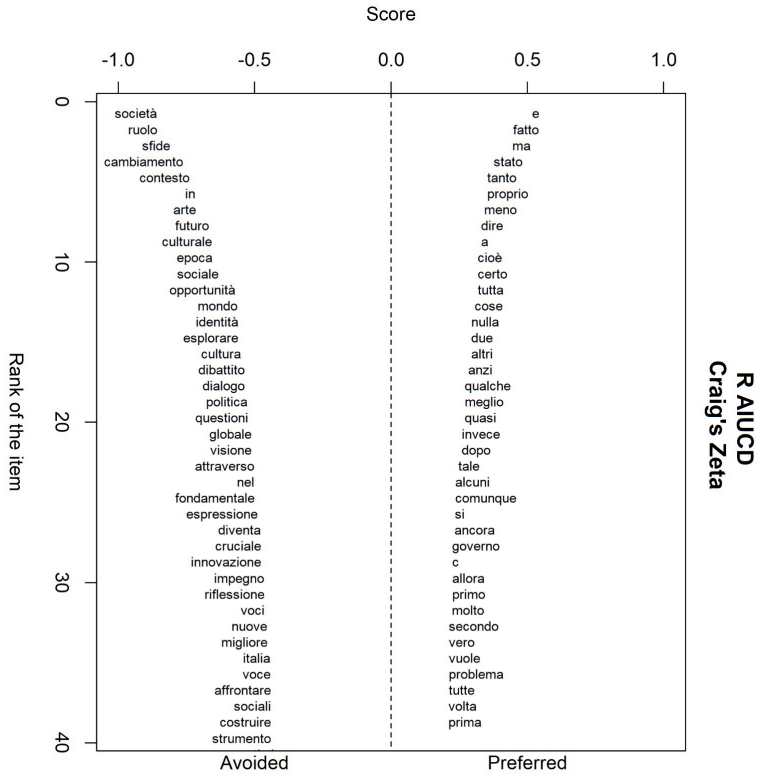


Figura 5. Zeta Analysis tra riviste e ChatGPT-4

6. CONCLUSIONI

Secondo quanto emerso nei paragrafi precedenti, risulta chiara la riconoscibilità dei testi *AI generated* rispetto a quelli scritti da umani. È altresì innegabile come l'emulazione di una specifica impostazione editoriale rappresenti una sfida ancora difficile da affrontare per i Large Language Models, dal momento che l'analisi stilometrica delinea una distanza tra i gruppi di testi. Ciononostante, lo studio potrebbe rivelarsi utile per futuri sviluppi, se consideriamo il processo di indagine adottato. Il lavoro su insiemi di testi, scritti da autori differenti ma pubblicati nella stessa sede, uno sguardo rivolto alla complessità, inevitabilmente variegata, di una rivista culturale e il tentativo di riconoscere una strategia editoriale ad ampio raggio potrebbero dischiudere nuovi orizzonti di indagine sulle pratiche di scrittura e di lettura, nonché sulle specifiche modalità di fruizione dei prodotti culturali in determinati periodi storici. Inoltre, il confronto tra testi realizzati da umani e da LLM idealmente finalizzati alla pubblicazione su una sede – il giornale letterario appunto – potrebbe portare a percorsi di approfondimento diversi intorno alle modalità di addestramento degli LLM stessi, alle loro attuali capacità generative e all'insieme di tecniche adottate sul loro sviluppo: individuare e riprodurre una strategia editoriale, d'altro canto, rappresenta una sfida ancora complessa e un terreno di studio da sondare ed esplorare prima di tutto nel campo dei *cultural studies*. In questo senso, potrebbero venire in soccorso tecniche di analisi computazionale, capaci di riconoscere tratti dirimenti tra le diverse posizioni editoriali attraverso uno sguardo d'insieme. Ecco perché è auspicabile che l'analisi stilometrica si rivolga in futuro, come solo in parte ha già fatto, anche a miscellanee che vadano dalle riviste alle raccolte collettive di opere di diversi autori e si affianchi a un approfondimento parallelo della capacità imitativa degli LLM rispetto a insiemi complessi di testi: un simile approccio potrebbe, infatti, condurre a importanti novità in quei vasti campi di indagine che sono la storia della lettura e la teoria della ricezione.

7. RINGRAZIAMENTI

Il progetto di Marco De Cristofaro è finanziato dall'Unione Europea nell'ambito del programma Horizon 2020 – Marie Skłodowska Curie – grant agreement No 101034383.

BIBLIOGRAFIA

[1] Balestrini, Nanni. *Quindici: una rivista e il Sessantotto*. Milano: Feltrinelli, 2008.

[2] Baranelli, Luca, e Grazia Cherchi. *Quaderni piacentini 1962-1968*. Milano: Gulliver, 1977.

[3] Bourdieu, Pierre. *Les Règles de l'art. Genèse et structure du champ littéraire*. Paris: Editions du Seuil, 1992.

[4] Burrows, John. «All the Way Through: Testing for Authorship in Different Frequency Strata». *Literary and Linguistic Computing* 22, fasc. 1 (2006): 27-47. https://doi.org/10.1093/llc/fqi067.

[5] Burrows, John. «"Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship». *Literary and Linguistic Computing* 17, fasc. 3 (2002): 267-287.

[6] Eder, Maciej. «Does Size Matter? Authorship Attribution, Small Samples, Big Problem». *Digital Scholarship in the Humanities* 30, fasc. 2 (2013): 167-182. https://doi.org/10.1093/llc/fqt066.

[7] Eder, Maciej, Jan Rybicki, e Mike Kestemont. «Stylometry with R: A Package for Computational Text Analysis». *R Journal* 8, fasc. 1 (2016): 107. https://doi.org/10.32614/rj-2016-007.

[8] Fofi, Goffredo, e Vincenzo Giacopini. *Prima e dopo il '68*. Roma: minimum fax, 1998.

[9] Guerriero, Stefano. «Salotto, laboratorio, dipartimento: la rivista come istituzione letteraria nel secondo Novecento, da "Aretusa" a «Linea d'Ombra». In *La letteratura italiana del Secondo Novecento fuori d'Italia: ricezione e immaginario (1945-1989)*, a cura di Alejandro Patat e Brigitte Poitrenaud Lamesi. Bruxelles: Peter Lang, 2021.

[10] Haaf, Susanne, Frank Wiegand, e Alexander Geyken. «Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text». *Journal of the Text Encoding Initiative* 4 (2013). https://doi.org/10.4000/jtei.739.

[11] Holley, Rose. «How Good Can It Get?» *D-Lib Magazine* 15, fasc. 3/4 (2009). https://doi.org/10.1045/march2009-holley.

[12] Kichuk, Diana. «Loose, Falling Characters and Sentences: The Persistence of the OCR Problem in Digital Repository E-books». *Libraries and the Academy* 1 (2015): 59-91. https://doi.org/10.1353/pla.2015.0005.

[13] Köbis, Nils, e Luca D. Mossink. «Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry». *Computers in Human Behavior* 114 (2021): 106-553.

[14] Muraca, Giuseppe. «Cronistoria dei 'Quaderni piacentini'». In *Da "Il Politecnico" a "Linea D'ombra"*. Poggibonsi: Lalli, 1990.

[15] Pontremoli, Giacomo. *I "Piacentini". Storia di una rivista (1962-1980)*. Roma: Edizioni dell'asino, 2017.

[16] Rebora, Simone. «GPT-3 vs. Delta. Applying stylometry to large language models». In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 292-297, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[17] Van der Lee, Chris, Albert Gatt, Emiel Van Miltenburg, e Emiel Krahmer. «Human evaluation of automatically generated text: Current trends and best practice guidelines». *Computer Speech & Language* 67 (2021): 101151. https://doi.org/10.1016/j.csl.2020.101151.

[18] Ward, Joe H. «Hierarchical Grouping to Optimize an Objective Function». *Journal of the American Statistical Association* 58, fasc. 301 (1963): 236-244. https://doi.org/10.1080/01621459.1963.10500845.

C’è un testo in questa chat?
Intelligenza artificiale e cooperazione interpretativa

Daniele Silvi
Università degli Studi di Roma ‘Tor Vergata’, Italia - d.silvi@me.co

ABSTRACT
Ispirato agli esperimenti di lettura di Stanley Fish, questo intervento vuole riproporli con l’ausilio dell’intelligenza artificiale e del chatbot ChatGPT. Nel suo *Is there a text in this class*, [2] il teorico americano fa credere ad alcuni studenti universitari che un elenco di nomi scritti sulla lavagna sia una poesia e – conseguentemente – ne chiede l’interpretazione. In questo poster descriverò l’esperimento fatto chiedendo a ChatGPT 4 di creare alcuni autori fittizi e di generare brani di loro ipotetiche opere. Presenterò inoltre i risultati di un sondaggio, condotto tra gli studenti universitari della mia Macroarea, per valutare l’attuale validità della teoria di Fish e del Pragmatismo, mettendola di più sotto stress per mezzo dell’Intelligenza artificiale. La tesi è che l’Intelligenza artificiale sia in grado di generare testi letterari non solo con le sue abilità ‘performative’ o ‘creative’ ma soprattutto in virtù della cooperazione interpretativa con il lettore.

PAROLE CHIAVE
ChatGPT; pragmatismo; teoria della letteratura; intelligenza artificiale.

1. INTRODUZIONE
Stanley Fish, studioso americano della seconda metà del ‘900, fu docente di Teoria della letteratura presso la Johns Hopkins University e la Duke University ma nel 1971 teneva due corsi per il Linguistic Institute of America e il Dipartimento di Inglese della State University of New York, a Buffalo. Le lezioni si svolgevano al mattino nella stessa aula. Alle 9.30 entrava un gruppo di studenti che seguiva il corso di linguistica e critica letteraria, a seguire invece si davano il cambio gli studenti i cui interessi erano esclusivamente di ordine letterario e di fatto si limitavano alla poesia religiosa inglese del XVII secolo. Questi studenti imparavano come identificare i simboli cristiani, come riconoscere le strutture tipologiche e come, muovendo dall’osservazione di tali simboli e strutture, giungere a descrivere un’intenzione poetica. In una mattina estiva di lezione, capitò che sulla lavagna era rimasto scritto un elenco di nomi di linguisti, che indicava un compito assegnato al primo gruppo. Quando gli studenti del secondo gruppo furono entrati, Fish disse loro che ciò che vedevano era una poesia religiosa del genere che stavano studiando, e chiese di darne un’interpretazione. Il loro comportamento rispose subito a un modello che era più o meno prevedibile, secondo gli intenti di Fish. Tutti si applicarono, fornendo spiegazioni coerenti e convincenti e nessuno mise assolutamente in dubbio che quella potesse essere davvero una poesia piuttosto che un elenco di nomi propri di persona. Il gruppo di alunni tentò di fornire delle interpretazioni, alle volte anche geniali, di quanto scritto sulla lavagna e così Fish giunse alla conclusione che presentando un qualunque testo come poesia ad un gruppo di persone che effettivamente si aspettavano di leggere una poesia, produceva l’effetto desiderato: esse vedevano una poesia. Ecco, quindi, che in qualche modo l’esperimento dimostra come le nostre supposizioni possano influenzare il pensiero al punto da creare una realtà testuale, anche laddove non ce ne è effettivamente una.
In questo intervento riproporrò l’esperimento di Fish con l’ausilio dell’Intelligenza artificiale, sottolineandone proprio l’aspetto di istanza di una moltitudine di intelligenze (e non quindi come una intelligenza generica) come quelle descritte dalla teoria di Gardner sulle intelligenze multiple [4].
Ho utilizzato ChatGPT 4, poiché il Large Language Model (LLM), il modello linguistico su cui si basa, è un esemplare della classe dei “transformer model” [7], modelli di reti neurali artificiali che tengono traccia delle relazioni all’interno delle sequenze di parole che vengono loro fornite come input, al fine di generare la prosecuzione linguistica più coerente: una tecnologia avanzata e innovativa nell’ambito dell’intelligenza artificiale e del machine learning che utilizza una struttura denominata appunto “trasformatore” per processare e comprendere i dati in modo più efficace rispetto ai metodi precedentemente utilizzati. La novità che ha portato questo metodo è basata sull’apprendimento automatico [3] che viene effettuato dalla macchina: andando a vedere le relazioni che intercorrono tra le parole, il modello riesce a produrre testi, anche di dimensioni notevoli.
Da questo punto di vista l’intervento che propongo si pone dialetticamente tra la testualità e l’organizzazione della conoscenza, proponendo una riflessione teorica e metodologica sulla percezione socio-culturale dei testi e della letterarietà [5].

2. APPROCCIO E METODOLOGIA

Per condurre l'esperimento, oggetto di questo paper, ho chiesto a ChatGPT di creare degli autori fittizi di diverse epoche e diversi stili. La prima richiesta è stata formulata come segue: *Puoi generare un autore letterario fittizio che abbia uno stile epico, romantico e idealizzato, che richiama l'immaginario e le convenzioni letterarie dell'epoca in cui sono ambientate le storie cavalleresche?*

La risposta è stata la nascita di Alessandro di Montalba, che ChatGPT così ci descrive: *Alessandro di Montalba è un poeta immaginario del XV secolo, appartenente alla corte di un nobile signore. Le sue opere spaziano tra le epopee cavalleresche e le liriche amorose, riflettendo il suo interesse per l'avventura e l'amore cortese.*

Alessandro, noto per la sua prosa raffinata e la maestria nell'uso di metafore elaborate, ha una predilezione per la descrizione dettagliata degli eroi e delle eroine delle sue storie. La sua narrativa è intrisa di un sentimento romantico, dove le gesta eroiche si intrecciano con le sottili sfumature delle relazioni umane.

La sua opera più celebre, forse, è un'epica cavalleresca intitolata Le gesta di Montalba*, un'opera maestosa che narra le imprese di un eroe immaginario alla corte di Carlo Magno. Nel contempo, le sue poesie d'amore, raccolte sotto il titolo* Cantari d'Innamoramento*, riflettono la sua sensibilità verso i temi della passione e della nostalgia.*

Alessandro di Montalba è ricordato come un maestro nel creare mondi immaginari intrisi di nobiltà e ideali cavallereschi, contribuendo così a plasmare la visione romantica della sua epoca.

Parallelamente ho chiesto al chatbot di generare anche uno stralcio del suo poema più noto, che avrei poi utilizzato per l'esperimento con gli studenti (e che descriverò nel prossimo paragrafo). Infine ho chiesto di generare un autore fittizio che scrivesse in latino (ed è nato Lucius Serenus, autore del I sec. D.C.) ed uno che componesse poesie in lingua inglese (ed è nata Seraphina Nightshade, poetessa inglese del primo XIX secolo). Per ognuno di questi autori è stato generato uno stralcio di testo e poi questi testi sono stati organizzati in un questionario che ho proposto ad un campione di 50 studenti (equamente divisi tra uomini e donne) del corso di laurea in Lettere dell'Università di Roma Tor Vergata.

3. ESPERIMENTO E RISULTATI

Gli studenti interpellati si sono mostrati disponibili e molto interessati all'argomento proposto, con una conseguente desiderio di conoscere le risposte esatte del test; la prima domanda (vd. Fig. 1) alla quale dovevano rispondere apparteneva alla collocazione temporale del testo che era stato da me falsamente proposto come generato da un'intelligenza artificiale. In realtà si trattava di una parte della *Secchia rapita* di Alessandro Tassoni, noto poema eroicomico del '600.

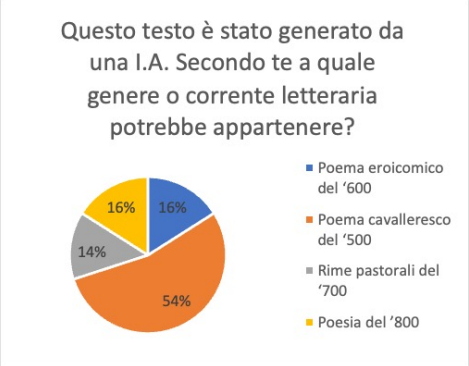


Figura 1

La seconda domanda era stata formulata con le stesse caratteristiche della prima, quindi anche qui, qualsiasi risposta data risultava essere sbagliata; tuttavia, possiamo notare (vd. Fig. 2) che gli studenti dando una prima occhiata veloce al testo, e individuando la parola: Orlando, immediatamente lo associavano a Ludovico Ariosto e, quindi, al suo poema cavalleresco: l'*Orlando furioso*, andando subito a cercarlo tra le risposte senza nemmeno fare caso alle altre.

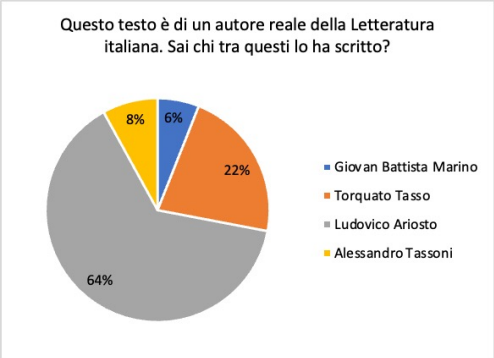


Figura 2

Il testo della terza domanda è stato invece generato artificialmente, come opera del già citato pseudo-Catullo, dal nome di Lucius Serenus. Gli studenti, appena leggevano il testo si trovavano però già in maggiore imbarazzo, forse perché non molti di loro appartenevano al curriculum classico degli studi e non riconoscevano tracce catulliane nei versi. Il risultato è stato un equo bilanciamento delle risposte, quasi una distribuzione normale (vd. Fig. 3).

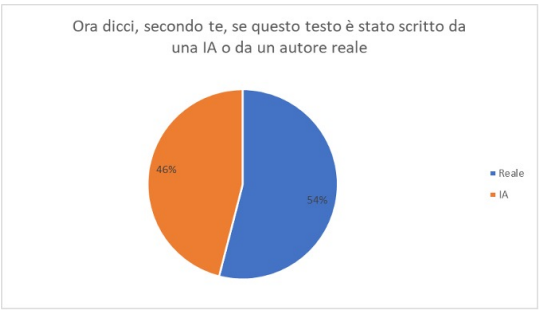


Figura 3

Questo esempio è particolarmente illuminante per la mia tesi, poiché mostra che quando iniziano a mancare i riferimenti letterari già conosciuti dal lettore (che chiamerò 'riferimenti interni') allora oscilla anche la sua capacità interpretativa. In sostanza il lettore sembra agire in maniera inversa rispetto all'intelligenza artificiale: egli utilizza le sue conoscenze pregresse per interpretare il testo che non conosce ma quando questi strumenti si rivelano insufficienti o inadatti allora il testo non è più tale, diventa un'invenzione della macchina [1].
Stesso discorso per l'ultima domanda che proponeva la scelta tra una poesia del poeta inglese William Blake e un'altra, generata da ChatGPT ma con lo stile del poeta sopracitato. Anche qui gli studenti interpellati, dopo aver letto entrambe le versioni proposte si sono distinti in due categorie: una di esse conosceva la poesia proposta e ha risposto con certezza, l'altra invece, ha optato per una risposta casuale, spaccando ancora una volta in due metà la distribuzione delle risposte (vd. Fig. 4).

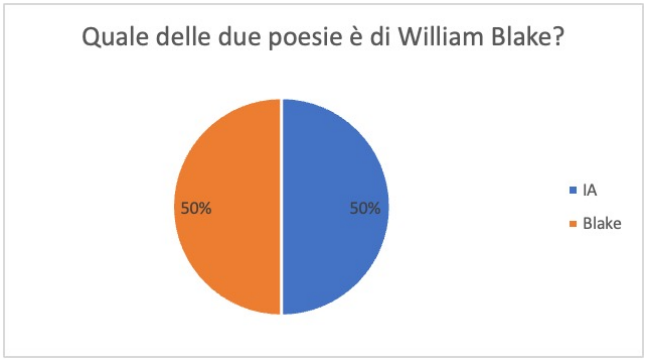


Figura 4

4. CONCLUSIONI

La mia ricerca vuole far riflettere su l'effettiva potenza di creazione dei testi da parte delle neonate intelligenze artificiali. Secondo lo studio condotto finora, la cooperazione interpretativa e l'apporto cognitivo umano sono ancora un ingrediente fondamentale della creazione letteraria, secondo l'approccio della ricezione. In sostanza ChatGPT è in grado di generare e simulare perfettamente degli autori fittizi, attribuendogli uno stile e delle opere ma è ancora l'uomo a vedere in certi testi la letterarietà o l'assenza di essa, in virtù del suo pregresso, della sua collocazione spazio-temporale e della concrezione critica delle sue letture precedenti [6]. Mi riservo di approfondire questa traccia di ricerca – in un futuro paper – ampliando l'indagine per individuare, ad esempio, la capacità degli studenti di riconoscere le strutture tipiche della lirica e come esse cambiano in base al genere o alla diacronia.

BIBLIOGRAFIA

[1] Benanti, Paolo. *Human in the loop. Decisioni umane e intelligenze artificiali*. Milano: Mondadori Università, 2022.

[2] Fish, Stanley. C'è un testo in questa classe? L'interpretazione nella critica letteraria e nell'insegnamento. Torino: Einaudi, 1987.

[3] Gardner, Howard. Educazione e sviluppo della mente. Intelligenze multiple e apprendimento. Roma: Erickson, 2005.

[4] Gardner, Howard. *Frames of mind: The theory of multiple intelligences*. New York: Basic Books, 2011.

[5] Lee, Kai-Fu, e Chen Qiufan. *Ai 2041. Scenari dal futuro dell'intelligenza artificiale*. Roma: Luiss University Press, 2023.

[6] Mitchell, Melanie. *L'intelligenza artificiale. Una guida per esseri umani pensanti*. Torino: Einaudi, 2022.

[7] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aiden N. Gomez, Lukasz Kaiser, e Illia Polosukhin. «Attention is all you need». *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2017, 1-11.

Genere e geopolitica nelle Discipline Umanistiche Digitali in Italia. Le Conferenze AIUCD (2012-2023)

Selenia Anastasi
Università degli Studi di Genova, Italia - selenia.anastasi@edu.unige.it

ABSTRACT
Questo studio indaga la diversità di genere e geopolitica nelle Digital Humanities (DH) in Italia, in particolare nel contesto dall'Associazione Italiana per l'Informatica Umanistica e la Cultura Digitale (AIUCD), su esempio di analisi precedenti sulle conferenze ADHO [3]. A partire da un corpus di atti di convegno pubblicati da AIUCD dal 2012 al 2023, l'analisi si concentra sulla rappresentazione di genere e sul divario Nord-Sud nel contesto istituzionale legato alla ricerca in DH. Questo lavoro preliminare mostra come, dal punto di vista della rappresentazione di genere, il quadro generale veda un lento miglioramento verso un maggiore coinvolgimento delle donne, mentre dal punto di vista geopolitico la comunità delle DH italiana è ancora accentrata su singoli e limitati poli d'eccellenza, prevalentemente nelle istituzioni tradizionalmente più legate all'Umanistica Digitale, ed economicamente più fiorenti del Settentrione. Lo studio invita a un esame critico e a una riflessione più generale sulla direzione futura delle DH, sottolineando inoltre l'importanza di politiche che promuovano attivamente una maggiore inclusività nel settore.

PAROLE CHIAVE
Genere e DH; inclusività; geopolitica; analisi quantitative; analisi dei networks.

1. INTRODUZIONE
In relazione al ricco e differenziato campo d'indagine delle Digital Humanities, le riflessioni sulla diversità e la rappresentazione, intesi nei loro sensi più ampi, ha una storia relativamente giovane. A partire dalle critiche di Liu [5] sulla carenza di approcci critici alle discipline umanistiche digitali, il discorso si è allargato nel tempo includendo considerazioni generali sul rapporto tra tecnologie, oppressione e giustizia sociale. Il culmine del dibattito è stato raggiunto nel 2019 con la pubblicazione della raccolta *Bodies of Information: Intersectional Feminism and the Digital Humanities* [6], una raccolta di contributi volti a indagare come le pratiche più diffuse all'interno delle DH tendano a rafforzare, e nel migliore dei casi a ignorare, disparità esistenti da prospettive che vanno dalle egemonie linguistiche a quelle di genere e geopolitiche. Il riconoscimento di questi meccanismi ha, d'altro canto, favorito lo sviluppo in positivo di un nuovo indirizzo di ricerca noto come Digital Humanities critiche [1], che intreccia considerazioni su razza, genere, sessualità, capitalismo delle infrastrutture e geopolitica della conoscenza a indagini sullo statuto della teoria e delle prassi nelle discipline umanistiche digitali [4]. Su questo fronte, è possibile tracciare approssimativamente tre approcci metodologici prevalenti. Il primo agisce al livello del contenuto e si pone come obiettivo quello di comprendere se, attraverso le piattaforme, le infrastrutture e gli strumenti digitali, le imprese DH stanno, volontariamente o meno, veicolando contenuti stereotipici o tesi a oscurare sistematicamente certe classi sociali [8]. Il secondo livello di indagine è analitico, e agisce identificando criticamente il modo in cui le tecnologie incoraggiano, supportano e accrescono le disparità tra generi e tra classi sociali, con la produzione di appropriati report e analisi quantitative a partire da dati concreti [2, 3]. Il terzo e ultimo livello agisce sulla struttura, ripensando i design, i modelli educativi, e le infrastrutture dal punto di vista della loro progettazione[1].
In questa cornice di ricerca più vicina ai Cultural Studies, e su impulso di ambiziosi e innovativi progetti[2] [8] volti a incorporare i principi formulati dalle epistemologie femministe alle architetture digitali e alle nuove tecnologie, le Digital Humanities Femministe (FDH) si sono affermate più recentemente e in modo autonomo, come naturale conseguenza di questo rinnovato sguardo. Sottolineando l'importanza di un approccio materialista ai dati e di una riflessione generale sui poteri coinvolti nella produzione, organizzazione e disseminazione della conoscenza, le FDH hanno dato impulso alla nascita di numerosi progetti con obiettivi e metodologie innovativi. Tra questi, vale la pena citare il lavoro sugli archivi digitali dell'Orlando Project e del Woman Writers Online (WWO), il più recente Full Stack Feminism in Digital Humanities[3], che coniuga esperienza didattica ed expertise tecnologica al fine di sviluppare un kit di strumenti

[1] https://fullstackfeminismdh.pubpub.org/
[2] http://www.arts.ualberta.ca/orlando/
[3] Vd. nota 1.

interoperabili (su tre livelli: archivi, codice, accesso), e alcune interessanti ricerche quantitative che hanno evidenziato la propensione delle DH a promuovere discorsi patriarcali o egemonici in ambito accademico [3].
Questo studio si posiziona all'interno del secondo livello di analisi, a partire cioè da un approccio descrittivo e analitico. Sebbene parziali, i risultati ottenuti offrono una fondamentale base di partenza per future indagini più dettagliate sulla rappresentazione della diversità nelle discipline umanistiche digitali in Italia.
Le DH rappresentano oggi un'interfaccia cruciale tra le metodologie tradizionali delle scienze umane e le nuove tecnologie digitali, promuovendo l'evoluzione della ricerca e delle pratiche accademiche. Tuttavia, resta ancora da indagare empiricamente il possibile impatto della ricerca in DH sulla distribuzione dei generi all'interno del contesto accademico umanistico, tradizionalmente caratterizzato da una forte presenza femminile – sebbene restino ancora esigui i casi di donne in posizioni accademiche di prestigio e in ruoli di leadership anche in questo campo. Infatti, da una parte l'emergere delle DH ha senza dubbio aperto nuove opportunità di ricerca rispetto alle metodologie accademiche tradizionali, consentendo una partecipazione più equa e diversificata e incoraggiando una cultura collaborativa interdisciplinare. D'altra parte, l'orientamento delle DH verso le grandi iniziative si traduce spesso in un'accentuata propensione verso soluzioni tecnologiche fini a sé stesse, e realizzate in collaborazione con entità istituzionali di vasta scala meno sensibili ad accogliere le prospettive provenienti da contesti geopolitici periferici.
Questo approccio, dunque, non offre soltanto uno sguardo generale sulle tendenze attuali, ma fornisce anche una base solida per l'elaborazione di interventi mirati volti a creare un ambiente accademico più inclusivo e rappresentativo.

2. IL CONTESTO GLOBALE: LE CONFERENZE ADHO

A motivare la presa di posizione femminista e intersezionale da parte delle studiose impegnate nelle DH si situano diverse iniziative che hanno evidenziato squilibri significativi nella rappresentazione della diversità a partire proprio dai luoghi di riferimento per la comunità scientifica. L'indagine di [3] sulla rappresentazione della diversità nelle conferenze ADHO, ha dato visibilità e forma concreta a una tendenza fino ad allora solamente percepita – a riprova del fatto che approcci critici alla cultura e studi quantitativi possono e devono convivere fruttuosamente. Il report raccoglieva i dati delle pubblicazioni ADHO tra il 2000 e il 2015 e riportava statistiche in materia di genere, affiliazioni, e topic delle proposte. Tra i risultati più interessanti, il report segnalava che, nonostante il risalto dato ad alcune studiose, le donne sono generalmente sottorappresentate negli incontri più importanti del settore (il 32.7% su un totale di 3.239 autori/trici erano donne, mentre solo il 29% sul totale ha presentato un contributo come singola autrice e il 33.4% ha co-autorato contributi).
Su esempio di questo lavoro, e con lo scopo esplicito di colmarne alcune carenze, nel 2015 Bordalejo [2] pubblica un sondaggio rivolto alla comunità ADHO. In questo studio, la diversità delle conferenze ADHO appare ancora più omogenea dal punto di vista etnico rispetto a quando la si osserva attraverso la lente del genere, invitando quindi a riflettere sulla questione da una prospettiva intersezionale. Anche lo studio precedente di Eichmann-Kalwara aveva sottolineato la medesima urgenza, denunciando i bias impliciti associati alla lingua o all'identità di appartenenza che potrebbero aver influito sui livelli di accettazione di autori e autrici con nomi non statunitensi. Dalla pubblicazione di questi report, ADHO si è impegnata a più riprese in iniziative volte a promuovere una maggior trasparenza nei criteri di selezione e inclusione dei partecipanti e delle partecipanti alle conferenze annuali. Tra le tante iniziative vi è l'istituzione di una task force per promuovere politiche relative all'inclusione linguistica e culturale, l'introduzione di un comitato per l'inclusione intersezionale (dal 2022) e la creazione di un apposito codice di condotta. Tali proposte rappresentano un traguardo importante e il segno di una accresciuta sensibilità verso approcci più etici nelle pratiche della comunità scientifica delle DH a livello globale.

3. IL CONTESTO ITALIANO: LE CONFERENZE AIUCD

Nel contesto italiano delle DH, il lavoro di raccolta dati, monitoraggio e introspezione disciplinare è ancora agli inizi. Un tentativo recente è [7], sui contributi presentati tra il 2014 e il 2017 alle conferenze AIUCD e CLiC-it, in cui è stata applicata un'analisi multidimensionale delle collaborazioni tra autori e autrici e le loro pratiche citazionali, evidenziando i rapporti di reciproca influenza tra la comunità di Linguistica Computazionale italiana e quella di riferimento per l'Umanistica Digitale. Tuttavia, il contributo è orientato all'analisi dei contenuti e citazionale, tralasciando importanti considerazioni che riguardano la parità di genere in connessione alle pratiche citazionali e co-autoriali tra le due comunità. L'analisi dei network evidenzia inoltre esclusivamente i rapporti tra i singoli autori, il coinvolgimento di attori non-italiani agli eventi annuali, e alcune percentuali relative alle affiliazioni più rappresentative dei clusters, tralasciando riflessioni critiche sul divario Nord-Sud che emerge implicitamente dai risultati osservabili (le università più rappresentate da entrambe le comunità appartengono tutte ad atenei del Nord Italia).
Volendo integrare ulteriori dati utili al quadro generale sulla composizione demografica della comunità DH italiana, si è preso in considerazione anche il report AlmaLaurea sulla condizione dei laureati in LM-43 (Metodologie informatiche per

le discipline umanistiche)[4]. Malgrado una desolante carenza di dati relativi al genere della popolazione studentesca analizzata, il dato più interessante è quello che emerge dalle percentuali di migrazione degli/delle studenti/esse in tutta la fase dell'esperienza universitaria. Da quanto emerge dal report 2020/2021 (dati pre-pandemia), il 44,5% dei laureati in DH cambia regione per raggiungere la sede universitaria. I dati Almalaurea non forniscono indicazioni precise sulle modalità di spostamento e sulle origini/destinazioni, ma se si considerano le tendenze più generali che hanno investito negli ultimi anni le università del Meridione, non sembra difficile ipotizzare la direzione del flusso[5].

Come si evince da questo breve stato dell'arte, siamo ben lungi dall'avere un quadro completo di ciò che accade alla comunità delle DH italiane dal punto di vista della rappresentazione della diversità di genere e geopolitica. Lo studio qui proposto procede con una mappatura temporale della rappresentazione di genere e con l'identificazione della rete di istituzioni più influenti. Proponendosi di colmare questi *gap* analitici lo studio integra, inoltre, dati più recenti. Lo studio esclude dati relativi agli eventi CliC-IT, così come eventi concernenti altre comunità scientifiche potenzialmente affini, concentrando la propria attenzione esclusivamente sulla comunità DH più rappresentativa del quadro italiano delle DH *strictu sensu*.

4. COMPOSIZIONE DEL DATASET, METODOLOGIA E RISULTATI

Il dataset è stato collezionato a partire dai book of abstracts, dai programmi, e dagli atti di convegno pubblicati da AIUCD dal 2012 al 2023. I nomi degli/delle autori/trici sono stati estratti automaticamente utilizzando uno script in Python, e successivamente integrati manualmente con metadati relativi al genere (M/F)[6], al ruolo (autore/trice o keynote), all'affiliazione, se l'autore/trice è primo o unico autore/trice del contributo. Non avendo accesso a informazioni più approfondite sul luogo di provenienza degli autori/trici, per l'esame della diversità geopolitica si è scelto di prendere in considerazione solamente il dato relativo all'affiliazione degli/lle autori/trici all'epoca della pubblicazione. Come per il lavoro di [2, 3], siamo consapevoli che l'assenza di queste informazioni può restituire un quadro parziale della situazione. Tuttavia, riteniamo che per gli scopi di questo lavoro siano sufficienti. Successivamente alla raccolta dati sono state formulate quattro domande di ricerca principali: 1. Come cambia il coinvolgimento di autori e autrici in relazione al genere e in proporzione alla partecipazione globale nel tempo? 2. Come cambia nel tempo la quantità di primi autori e autrici in relazione al genere e in relazione al numero di pubblicazioni? 3. Come cambia nel tempo la quantità di primi autori e autrici in relazione al genere e in relazione al numero totale di autori coinvolti nelle pubblicazioni? 4. Quali affiliazioni sono più rappresentate e influenti nel contesto AIUCD?

Figura 1. Andamenti nel tempo relativi alla rappresentazione di genere (generale, autori/autrici singoli/e, e primi/e autori/trici (2012-2023)

[4] Report Almalaurea 202/2021 sui laureati in Metodologie Informatiche per gli Studi Umanistici: https://www.almalaurea.it/sites/almalaurea.it/files/docs/news/rapporto2021_almalaurea_approfondimento_digital-humanities.pdf.
[5] https://www.roars.it/perche-gli-atenei-del-sud-rischiano-di-scomparire
[6] L'autrice è consapevole che la scelta di rappresentare il genere secondo i canoni del binarismo è un limite, e non rispecchia il posizionamento politico di chi scrive.

Gli andamenti visualizzati in Fig. 1 sono normalizzati per tenere conto del totale dei/delle partecipanti/e, cresciuta complessivamente nel tempo. Nel linear chart numero 3 osserviamo le tendenze che rispondono alla prima domanda di ricerca. Nel 2012, le autrici rappresentavano circa il 30,95% sul totale. Nel 2013, la percentuale di autrici è aumentata al 42,50%, raggiungendo una situazione di quasi equilibrio nella rappresentazione di genere nel 2015 (49,40% per le autrici e 50,60% per gli autori). Tuttavia, nel 2016, la percentuale di autrici è nuovamente scesa al 34,91%. Tendenza che muterà solo nel 2021, con un lieve vantaggio delle autrici sugli autori (52,38% di autrici sul totale). L'anno successivo (2022), la percentuale di autrici è scesa drasticamente al 34,75%, risollevandosi nuovamente lo scorso anno (48,96% sul totale). Dal conteggio sono stati rimossi casi in cui un/a autore/trice abbia presentato e/o co-autorato più di un contributo nello stesso evento. Per quanto concerne la seconda domanda di ricerca, come si osserva nel linear chart numero 2, la rappresentazione delle prime autrici riscontra un incoraggiante incremento nel tempo. Anche in questo caso, tuttavia, solo in alcuni anni (2020 e 2013) si segnala un lieve vantaggio delle autrici sugli autori.
Molto più altalenante il quadro che riguarda la rappresentazione di genere in relazione agli autori e alle autrici singoli/e. Tuttavia, solo nel 2018 e nel 2023 si registra un significativo aumento per le autrici singole rispettivamente del 62% (2018) e del 60% (2023). Osservando il quadro generale, la rappresentazione delle donne come autrici di un contributo è aumentata di circa +18 punti percentuale.
Sul piano della diversità in relazione all'affiliazione, la Fig. 2 mostra la distribuzione geografica degli/delle autori/trici da Nord a Sud Italia. La colorazione dei cerchi sulle aree geografiche della cartina ci aiuta a identificare ad occhio le aree di maggior concentrazione dei contributi, con Lazio, Toscana, Emilia-Romagna e Veneto come le Regioni più rappresentate (blu e viola), mentre, nel complesso, si osserva una distribuzione piuttosto omogenea da Nord, al Centro e Sud Italia (con l'eccezione di Campania e Puglia per il Sud e di Piemonte, Lombardia e Trentino a Nord).

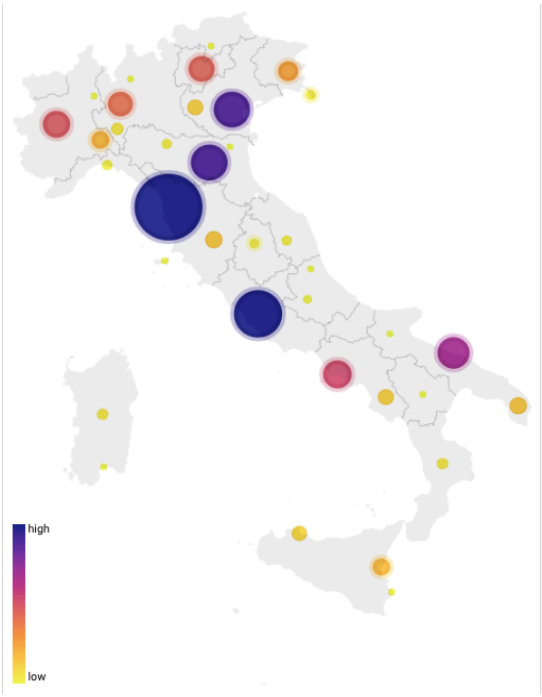


Figura 2. Distribuzione geografica affiliazioni al Nord, Centro e Sud Italia

Per cercare di illuminare più da vicino le aree di influenza in relazione alle affiliazioni degli autori e delle autrici, ci siamo avvalsi di una analisi dei networks, dove i nodi rappresentano le affiliazioni degli/delle autori/trici e gli archi una relazione di co-autorato nel contesto AIUCD (vd. Fig. 3). La visualizzazione dei networks è stata realizzata grazie al software open source Gephi[7].
Il grafo numero 2 in Fig. 3 rappresenta le affiliazioni con una *centralità di grado* elevata, dove i nodi di diametro maggiore corrispondono alle istituzioni che hanno il maggior numero di connessioni dirette con altri nodi del network. Queste affiliazioni sono importanti per la loro capacità di interagire direttamente con una vasta rete di altre entità. Inoltre, una centralità di grado elevata può indicare una posizione influente all'interno del network dovuta alle molteplici collaborazioni dirette.
Le affiliazioni con una *centralità di intermediazione* elevata (grafo numero 1, vd. Fig. 3) agiscono come snodi o intermediari nel flusso delle informazioni o delle collaborazioni all'interno del network. Queste affiliazioni possono controllare e influenzare la diffusione di idee, informazioni e risorse. Infine, le affiliazioni con una *centralità di vicinanza* elevata (grafo numero 3, vd. Fig. 3) sono quelle che, in media, possono diffondere informazioni in modo efficiente all'interno del network, raggiungendo più rapidamente altre entità. La loro posizione centrale nel network le rende strategicamente importanti per la rapida disseminazione di contenuti o per iniziative che richiedono un'ampia portata.

[7] https://gephi.org/

In tutti e tre i casi, in testa alle prime dieci affiliazioni con una centralità maggiore figurano la triade composta dall'Università di Bologna, l'Università e il CNR-ILC di Pisa e l'Università di Padova.

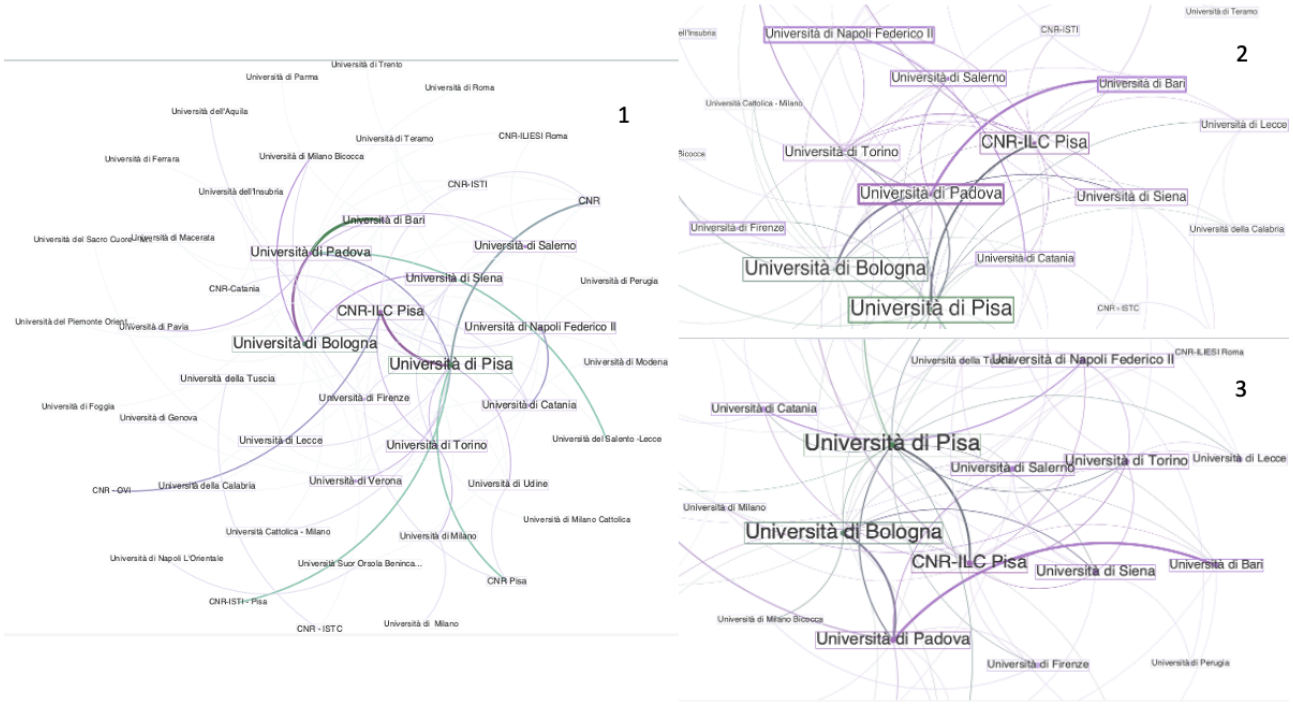


Figura 3. Networks affiliazioni autori/autrici AIUCD (2012-2023). Misure: centralità di intermediazione (1), centralità di grado (2), centralità di vicinanza (3)

In modo più interessante, le Università del Sud Italia (e in particolare le Università di Catania, di Bari, Napoli Federico II e Salerno) sembrano giocare un ruolo rilevante per la loro capacità di intermediazione all'interno del grafo.

5. CONCLUSIONI

Il lavoro di mappatura della rappresentazione della diversità all'interno di una comunità scientifica è cruciale per il progresso e l'integrità di qualsiasi disciplina, assicurando che rimanga rilevante, responsabile e sensibile ai cambiamenti sociali e ai progressi nel campo del sapere. Questo lavoro preliminare ha mostrato come, dal punto di vista del genere, il quadro generale veda un lento miglioramento verso un maggiore coinvolgimento delle donne, mentre dal punto di vista geopolitico la comunità delle DH italiana è ancora accentrata su singoli e limitati poli d'eccellenza, prevalentemente nelle istituzioni tradizionalmente più legate all'Umanistica Digitale ed economicamente più fiorenti del Settentrione. Studi futuri potrebbero approfondire le relazioni tra il genere e le affiliazioni di riferimento di autori e autrici, così come la percezione e l'interesse dei partecipanti alle conferenze in tema di diversità. Nonostante le numerose limitazioni di questo studio, riteniamo che i risultati possano favorire l'identificazione di potenziali aree di miglioramento, promuovere iniziative per incoraggiare un'adesione alla comunità realmente orizzontale, e stimolare un dialogo fruttuoso sul ruolo e il significato delle discipline umanistiche digitali nel panorama accademico nel suo insieme.

BIBLIOGRAFIA

[1] Berry, David M. «Critical digital humanities». In *The Bloomsbury Handbook to the Digital Humanities*, (a cura di) James Osullivan, Bloomsbury Publishing., 125-131. USA, 2022.

[2] Bordalejo, Barbara. «Minority Report: The Myth of Equality in Digital Humanities». In *Bodies of Information: Intersectional Feminism and the Digital Humanities*, (a cura di) Elizabeth Losh e Jacqueline Wernimont, 320-343. Univ of Minnesota, 2019.

[3] Eichmann-Kalwara, Nickoal, Jeana Jorgensen, e Scott B. Weingart. *Representation at Digital Humanities Conferences (2000-2015)*. JSTOR, 2018.

[4] Fiormonte, Domenico. «Digital humanities and the geopolitics of knowledge». *Digital Studies/Le Champ Numérique* 1, fasc. 7 (2017): 1-18.

[5] Liu, Alan. «Where is cultural criticism in the digital humanities?» *eScholarship, University of California*, 2012. https://escholarship.org/uc/item/2r66q4j0.

[6] Losh, Elizabeth, e Jacqueline Wernimont. *Bodies of information: Intersectional feminism and the digital humanities*. University of Minnesota Press, 2019.

[7] Sprugnoli, Rachele, Gabriella Pardelli, Federico Boschetti, e Riccardo Del Gratta. «Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale». *Umanistica Digitale* 5 (2019): 59-89.

[8] Wernimont, Jacqueline, e Julia Flanders. «Feminism in the age of digital archives: the women writers project». *Tulsa Studies in Women's Literature* 29, fasc. 2 (2010): 425-435.

Gli LLM come lettori modello artificiali

Fabio Ciotti
Università di Roma Tor Vergata, Italia - fabio.ciotti@uniroma2.it

ABSTRACT
Questo contributo esplora le potenzialità dei Large Language Models (LLM) nel campo della narratologia, e propone di considerali come lettori modello artificiali, riattualizzando il concetto della narratologia classica e in particolare dal lavoro di Umberto Eco. Attraverso un esame dello stato dell'arte e di esperimentazioni specifiche, il lavoro dimostra che gli LLM dispongono di competenze linguistiche e cognitive che li rendono capaci di comprensione narrativa, inferenza e generazione di ipotesi interpretative. Sebbene l'approccio sia esplorativo, i risultati preliminari suggeriscono che gli LLM potrebbero rappresentare un'innovazione paradigmatica per l'analisi letteraria e la comprensione testuale, promuovendo ulteriori indagini sistematiche sulle loro capacità.

PAROLE CHIAVE
Large Language Models; narratologia; comprensione narrativa; lettore modello; analisi letteraria computazionale.

1. INTRODUZIONE
La diffusione su vasta scala dei cosiddetti sistemi di Intelligenza Artificiale generativa, in particolare quelli linguistici o Large Language Models (LLM), ha rappresentato senza dubbio la maggiore innovazione nelle tecnologie digitali e computazionali di questo scorcio di secolo. L'impatto socioculturale di questa innovazione è ancora difficilmente valutabile, ma le capacità di questi sistemi nel produrre contenuti multimodali e nel gestire complesse interazioni linguistiche, a un livello di complessità ed efficacia ancora nemmeno immaginabile a inizio decennio, non possono essere sottovalutate. Il dibattito scientifico su tema è molto esteso e si articola su una vasta gamma di ambiti tematici di discussione (filosofico, psicologico, linguistico, informatico, ingegneristico, sociologico, economico…) e di posizioni che vanno da quelle dei detrattori più inflessibili a quelle degli apologeti più convinti, passando per un ampio spettro di posizioni più caute o interlocutorie. La ricostruzione di questo dibattito multipolare e multidisciplinare esula dai limiti di questo intervento. Mi concentrerò piuttosto su un sottoinsieme specialistico di queste discussioni, che riguarda un ambito specifico delle scienze della cultura, ovvero lo studio dei testi narrativi. Ho diverse remore nell'usare il temine tecnico narratologia, perché esso ha vaste implicazioni e molteplici accezioni, ma potremmo dire che questo paper intende esplorare alcune potenzialità dell'applicazione dei modelli linguistici generativi in ambito narratologico.
Questa indagine si pone in continuità e rappresenta una specificazione e un approfondimento sia teorico sia pratico, dei contenuti di [4] poi esteso in forma di articolo su [5]. Ma ha ancora un carattere esplorativo e non è uno studio esaustivo e su vasta scala corredato di valutazioni di accuratezza ed efficacia. D'altra parte, quando si tratta di LLM di nuova generazione l'efficacia e l'esaustività di una valutazione basata solo su benchmark quantitativi – ormai comune nei lavori in ambito di Machine Learning e NLP – è stata giustamente messa in dubbio [3, 18], e con buone ragioni. Sistemi complessi come GPT-4 e simili possono essere trattati sistemi intenzionali, e le loro capacità e competenze indagate analizzando i resoconti delle interazioni conversazionali con il modello durante l'esecuzione di un determinato compito.

2. STATO DELL'ARTE
L'esplorazione delle possibilità analitiche offerte dai modelli linguistici basati su word embedding e reti neurali allo studio della comunicazione testuale, e di quella letteraria in particolare, ha avuto inizio sin dalla diffusione dei primi modelli di embedding lessicale word2vec [17] e GloVe, come nel lavoro pionieristico di Heuser [11] sui testi del corpus ECCO-TCP. Successivamente al rilascio del primo modello a transformers BERT da parte di Google [6], grazie alla sua adattabilità mediante *fine tuning* a task di classificazione e sentiment analysis, le applicazioni di questi sistemi in contesti di ricerca umanistica si sono moltiplicate (si veda, ad esempio, l'elenco degli interventi presentati alle conferenze *Computational Humanities Research*[1]). L'elaborazione teorica e le applicazioni di Gavin in *Literary Mathematics* [10] rappresentano la sintesi teorica più recente e matura sull'analisi computazionale della testualità letteraria, anche se va detto che la monografia è stata scritta prima dell'introduzione di ChatGPT/GPT-4 e dei suoi successori.
L'esplorazione delle potenzialità di questi LLM di ultima generazione nell'analisi letteraria, in effetti, è in fase aurorale. Alcune ricerche esplorative sono state condotte da Underwood e i risultati preliminari pubblicati nel blog dello studioso. La prima [25] riguarda l'utilizzo dei modelli linguistici per misurare il passaggio del tempo narrativo nei testi di finzione.

[1] https://2023.computational-humanities-research.org

La seconda [24] si concentra sull'analisi della capacità di GPT-4 di prevedere gli sviluppi narrativi e sulla comparazione di queste previsioni con le reazioni intuitive dei lettori umani, offrendo così una prospettiva unica sull'intersezione tra intelligenza artificiale e studi empirici sulla recezione letteraria. Diversi paper sull'applicazione di LLM e/o *text embedding* sono stati presentati alla più recente edizione della già citata conferenza CHR [2, 7, 12, 14, 15, 21]. Infine, mette conto menzionare il meritorio lavoro di sviluppo e fine tuning di modelli dedicati alla ricerca umanistica di Langlais come il modello *Brahe* "an analytical LLM for multilingual literature fine-tuned from llama-13B. Given any text, Brahe will generate a list of potentially twenty annotations. Brahe is intended to be used by computational humanities project, similarly to BookNLP"[2].

3. NATURA E CAPACITÀ DEGLI LLM

Gli esempi citati nella sezione SOTA ci mostrano alcune delle possibili linee di indagine in questo vasto campo di possibilità analitiche, la cui esplorazione, assai promettente, è iniziata solo negli ultimi mesi. Tuttavia, per valutare adeguatamente se i metodi di analisi del testo narrativo basati sugli LLM possano rivelarsi una vera e propria svolta paradigmatica nell'ambito degli studi letterari computazionali, occorre anche un adeguato inquadramento teorico. Infatti, se la comprensione teorica, sia matematico-computazionale sia linguistico-letteraria, era essenziale per comprendere, applicare e valutare adeguatamente i metodi che la data science ha messo a disposizione degli analisti e teorici della cultura (dalla stilometria basata su *Most Frequent Words*, all'analisi probabilistica dei topic, alla network analysis), questa diventa ancora più fondamentale per le applicazioni degli LLM. Ciò è dovuto al fatto che, a differenza dei metodi precedenti, i modelli basati su reti neurali di grandi dimensioni hanno il problema di non essere perspicui: sebbene i processi matematici alla base del loro funzionamento siano ampiamente conosciuti, per ora le loro performance in fase di esecuzione non sono spiegabili meccanicisticamente, se non in casi estremamente semplificati.

In questa sede ci limiteremo a ricordare che, a un alto livello di astrazione, un LLM si può descrivere come un sistema in grado di predire su base probabilistica quale sia il token linguistico che segue una data sequenza di token (*prompt*), con la capacità di ripetere autonomamente il processo in modo autoregressivo fino alla produzione di un frammento di testo coerente. Tuttavia, questa descrizione di alto livello è troppo generale per dare conto del reale funzionamento e delle sorprendenti capacità acquisite da questi modelli. Per capire di cosa realmente stiamo parlando, la caratterizzazione matematica di alto livello deve essere sostituita da una descrizione di livello architetturale, che ci permetta di capire come funziona un LLM (modulo le varie ottimizzazioni e differenze di architettura di dettaglio che ogni sviluppatore di modelli può implementare, e di cui spesso non si ha documentazione, anche perché allo stato attuale i leader dello sviluppo innovativo in questo settore sono più aziende private che centri di ricerca universitari pubblici). A questo livello di descrizione, sappiamo che un LLM è una rete neurale che si basa sull'architettura a transformer e sul meccanismo del calcolo dell'attenzione; per una descrizione da un punto di vista umanistica su questi aspetti rimandiamo di nuovo a [5].

Ora, la cosa interessante è che questi modelli, addestrati a predire la parola più adeguata a proseguire una frase e poi specializzati nel preferire alcune strategie discorsive piuttosto che altre, hanno mostrato una serie di capacità emergenti, che non riguardano tanto e solo la loro conoscenza dichiarativa (ciò che sanno), spesso soggetta a errori o, come si usa dire, allucinazioni, ma un insieme di competenze linguistico-cognitive di alto livello (come tradurre in più lingue, produrre sommari e sinossi, effettuare processi di ragionamento abduttivo, esprimersi con diversi registri e socioletti, ecc.).

Il dibattito teorico e la ricerca sperimentale sulla reale sussistenza e sui limiti di queste capacità sono aperti e anche assai controversi, con lavori che certificano le incomprensioni e le fallacie degli LLM, a cui si contrappongono altri che ne dimostrano gli incredibili successi in vari ambiti di performance linguistiche e cognitive. Ovviamente, da un punto di vista teorico, la domanda è se questi modelli siano capaci o meno di 'vera' comprensione linguistica, di 'vera' capacità di ragionamento astratto, di coerenza di pianificazione e strategie, di creatività e innovazione. Anche se alcuni risultati sperimentali negativi – il più noto e rilevante dei quali è quello sulla "maledizione dell'inversione" [1] – rappresentano un solido argomento a favore di una posizione 'deflazionista' sugli LLM, e anche se i limiti di competenza semantica fattuale sono indiscutibili – da cui il consiglio di non affidarsi a un LLM come a un sostituto di un motore di ricerca generalista o verticale e a fonte di informazione architetturale come direbbe Roncaglia [22] –, sono altrettanto indiscutibili le loro capacità di elaborazione linguistica e, almeno in parte, di ragionamento. Anche senza fare alcuna assunzione sul fatto che la rappresentazione ed esecuzione del linguaggio degli LLM siano identiche a quelle umane, o che siano implementate fisicamente nello stesso modo in cui sono implementate in un cervello biologico, a parere di molti studiosi esse non sarebbero spiegabili se non ci fosse qualcosa di profondo relativamente al funzionamento del linguaggio che questi modelli riescono a catturare [19].

[2] https://huggingface.co/Pclanglais/Brahe

4. GLI LLM COME LETTORI MODELLO

Fatte queste premesse, veniamo al tema primario di questo paper, che riguarda la loro capacità di comprensione del testo narrativo. Alcune premesse: 1) non mi occupo qui della creazione narrativa da parte degli LLM. Su questo ovviamente c'è molta curiosità, qualche sperimentazione e un certo dibattito teorico, ma non è tema di discussione in questa sede anche perché competenza creativa e competenza interpretativa non si implicano reciprocamente; 2) non intendo qui assumere o supportare teorie forti circa la natura e le capacità cognitive generali degli LLM, ascrivere loro 'vere' credenze o 'genuine' proprietà semantiche, né sostenere che siamo di fronte a una vera *Artificial General Intelligence*, qualsiasi cosa voglia dire questa formula. Sulla base dell'osservazione empirica condotta in questi mesi e dell'analisi della letteratura scientifica, tuttavia, propongo di assumere le seguenti tesi:

(1) Gli LLM hanno una estesa competenza linguistica sia grammaticale sia semantica;
(2) Gli LLM posseggono una estesa e sistemica rappresentazione (non simbolica) dell'enciclopedia semantica di molti sistemi culturali, le cui manifestazioni linguistiche sono incluse nei loro *training set* [16]; tale rappresentazione è plurale e intersezionale, a causa della *superposizione* semantica di un LLM su insiemi arbitrari di parametri della sua rete neurale [9];
(3) Gli LLM hanno la capacità di comprendere almeno in parte implicature conversazionali e regole pragmatiche che governano l'uso del linguaggio da parte degli esseri umani;
(4) Gli LLM mostrano capacità inferenziali in contesti non-formali;

Nel dibattito narratologico di ambito semiotico del secolo scorso, ha avuto un ruolo molto importante la nozione di lettore modello [8], e sue molte varianti come quella di lettore implicito [13] o lettore virtuale [20]. Sarebbe troppo lungo qui ripercorrere le discussioni al riguardo e tracciare le differenze e le sovrapposizioni, peraltro ormai abbastanza fuori dai radar del dibattito teorico-letterario alla moda oggi, [cfr. 23]. Tra le varie formulazioni preferisco adottare quella proposta da Eco in *Lector in fabula* [8]:

> Per organizzare la propria strategia testuale un autore deve riferirsi a una serie di competenze […] che conferiscano contenuto alle espressioni che usa. Egli deve assumere che l'insieme di competenze a cui si riferisce sia lo stesso a cui si riferisce il proprio lettore. Pertanto, prevederà un Lettore Modello capace di cooperare all'attualizzazione testuale come egli, l'autore, pensava, e di muoversi interpretativamente così come egli si è mosso generativamente.

Il lettore modello secondo la teoria della cooperazione interpretativa di Eco deve avere una serie di competenze linguistiche testuali e meta-testuali:

(5) Competenza grammaticale
(6) Competenza semantico-enciclopedica
(7) Capacità di disambiguare gli impliciti
(8) Capacità di fare inferenze

La sovrapposizione o almeno prossimità concettuale delle competenze elencate nei punti (1)-(4) con quelle nei punti (5)-(8) mi porta ad avanzare la ipotesi seguente:

(9) i LLM possono essere una implementazione computazionale della nozione di lettore modello.

5. ALCUNE SPERIMENTAZIONI

La tesi che gli LLM siano dei lettori modello artificiali, per non essere considerata una mera metafora intellettuale, deve essere suffragata da evidenze sperimentali e deve avere un ruolo analitico ed esplicativo rilevante nella comprensione della comunicazione narrativa. Sulla competenza linguistica degli LLM esiste una vastissima serie di conferme, non ultima l'esperienza quotidiana di ogni utente di ChatGPT; in modo intuitivo possiamo assumere che gli LLM abbiano una vasta competenza enciclopedica che è delimitata dalla dimensione del training set usato nella fase di *self-supervised pretraining* e raffinata in quella di *Supervised Fine Tuning*. Si badi che questa competenza enciclopedica non va scambiata con la nozione epistemologica di conoscenza come insieme di credenze vere giustificate: come rilevava Eco l'enciclopedia semantica registra entità reali e fatti verificati insieme a finzioni e falsità, e il linguaggio può essere usato indifferentemente

per dire delle une e delle altre. Anche questo argomento andrebbe ulteriormente analizzato e approfondito ma qui preferisco puntare l'attenzione sulle tesi (3) e (4) avanzate nel paragrafo precedente. L'attribuzione agli LLM di tali competenze è molto controversa e su di essa si hanno evidenze sperimentali contraddittorie attestate dalla letteratura scientifica.
Come anticipato non ho condotto in questa fase una sperimentazione sistematica su ampia scala, preferendo in prima istanza un approccio conversazionale mediante l'interfaccia interattiva di ChatGPT, di cui fornisco in questa sede due esemplificazioni. Il primo esempio riguarda la capacità di GPT-4 di comprendere implicature conversazioni e presupposizioni, anche quando la decodifica richieda la conoscenza di strategie retoriche complesse come l'ironia. Nel primo prompt di questa interazione sono state indicate alcune istruzioni di base ed è stata fatta la richiesta di esplicitare la catena di ragionamento per ogni conclusione. Ecco il risultato:

User
Example (2)
Fabio this morning went to take his motorbike to go to school. He had woken up late and had to hurry not to arrive after the closing of doors at school. The weather was getting worse, and probably it would have rained very soon. He started the bike and drove to school, under the first raindrops. Halfway to school, the bike stopped working and the rain got a real downpour. He had to leave the motorbike there and walk to school, arriving late and wet. Later during lunch break he met Rita, his girlfriend, and had with her this conversation:

Rita: Hi Fabio, how are you doing?
Fabio: Oh well it's been really a lovely day!

Question: what is Fabio really meaning?

ChatGPT
Given the context and Fabio's response, it seems that Fabio's statement "Oh well it's been really a lovely day!" is likely sarcastic. His day has been problematic, with multiple issues including waking up late, his motorbike breaking down, and the rain, causing him to arrive at school late and wet. Here's my chain of thoughts:

[…]

Fabio's statement contradicts the situations described earlier. He uses positive language ("really a lovely day") to describe a series of negative experiences.
Sarcasm as a Conversational Tool (Relation & Quality)

Sarcasm often involves saying something but meaning the opposite, usually to convey frustration, disappointment, or other emotions. In this context, sarcasm serves to express Fabio's feelings about the day more vividly.
Contextual Information (Relation)

The detailed context provided about Fabio's unfortunate events sets the stage for interpreting his statement. Without this context, the sarcasm in his statement might not be as apparent.
Interpersonal Dynamics

Given that Rita is Fabio's girlfriend, there is likely a level of understanding and familiarity between them that allows for sarcasm and indirect expression of feelings. Fabio might expect Rita to understand his real feelings behind the sarcastic remark.
Conclusion:
Fabio is likely using sarcasm to express his frustration and disappointment about the day's events. His usage of the term "lovely day" is contradictory to the series of unfortunate events he experienced, and given the context, it seems that he is indirectly communicating his real feelings—of frustration and disappointment—to Rita.

Il secondo esempio invece è più attinente alle questioni teoriche legate alla nozione di lettore modello nel testo narrativo poiché cerca di valutare la capacità di GPT-4 di effettuare quelle che Eco chiama *passeggiate inferenziali*, un meccanismo centrale nel processo della cooperazione interpretativa: data una sequenza narrativa che porta a un punto di svolta nell'intreccio, la capacità di elaborare diversi possibili sviluppi narrativi, di compiere "il gesto libero e disinvolto con cui il lettore si sottrae alla tirannia del testo – e al suo fascino – per andare a ritrovare esiti possibili nel repertorio del già detto" [8: 118].

User
You act as a model reader. I will give a passage of narrative. Read it and provide two or three hypothesis on what will happen next.

TEXT TO ANALYZE
S'era levata la giacca per portar dentro la catapecchia alcuni covoni, una ventina, che dopo la mietitura il contadino non aveva ancora trasportato nel suo podere di là e aveva lasciato qua all'aperto. Il cielo, durante la notte, s'era incavernato, e la pioggia pareva imminente. Abituato a non far mai nulla, per quella fatica insolita e per quella sciocca previdenza, che poi del resto non spettava neanche a lui perché quei covoni di grano appartenevano come tutto il resto al contadino, s'era tanto stancato, che quando fu per trovar posto dentro la catapecchia, già tutta stipata, all'ultimo covone, non ne potè più, lasciò quel covone davanti la porta, e sedette per riposarsi un po'.

A capo chino, con le braccia appoggiate alle gambe discoste, lasciò penzolare tra esse le mani. E a un certo punto ecco che si vide uscire dalle maniche della camicia su quelle mani penzoloni le formiche, le formiche che dunque sotto la camicia gli passeggiavano sul corpo come a casa loro. Ah, perciò forse la notte lui non poteva più dormire e tutti i pensieri e i rimorsi lo riassalivano. S'infuriò e decise lì per lì di sterminarle. Il formicaio era a due passi dalla porta. Dargli fuoco.

Come non pensò al vento?

ChatGPT
La narrazione presenta un contesto in cui il protagonista è coinvolto in un'attività insolita per lui, la movimentazione di covoni di grano, all'interno di un ambiente che suggerisce un'imminente pioggia. L'episodio delle formiche che emergono dalle maniche della sua camicia rappresenta un elemento di disturbo che scatena in lui una reazione impulsiva.

Ipotesi di Continuazione basata sull'Impulsività del Personaggio: Il protagonista, già dimostratosi impulsivo e irritabile, potrebbe agire in modo avventato e sconsiderato per distruggere il formicaio. Questa azione potrebbe avere conseguenze inattese, come la diffusione del fuoco a causa del vento menzionato nel testo, portando a un incendio non controllato. Tale sviluppo potrebbe introdurre un conflitto maggiore nel racconto, dove il personaggio deve affrontare le ripercussioni delle sue azioni impulsive.

Ipotesi basata sul Simbolismo delle Formiche: Le formiche potrebbero essere interpretate come un simbolo dei pensieri e rimorsi del personaggio. La sua decisione di "sterminarle" potrebbe indicare un tentativo di affrontare o sopprimere questi pensieri. Il fallimento o la difficoltà nell'eliminare le formiche potrebbe simboleggiare l'incapacità del personaggio di affrontare i suoi problemi interni, portando a una maggiore introspezione o a un ulteriore declino psicologico.

Ipotesi basata sull'Elemento Ambientale: La menzione del cielo "incavernato" e della pioggia imminente suggerisce che gli elementi ambientali potrebbero giocare un ruolo significativo nella continuazione della storia. La pioggia potrebbe iniziare a cadere nel momento in cui il protagonista tenta di distruggere il formicaio, possibilmente ostacolando il suo piano e introducendo nuove sfide o cambiamenti nella narrazione, come la necessità di trovare riparo o la conseguenza di lasciare il covone di grano fuori dalla catapecchia.

Queste ipotesi sono basate sull'analisi testuale e su elementi narrativi ricorrenti in letteratura. Tuttavia, esse rimangono congetture e interpretazioni soggettive del testo fornito.

Anche in questo caso, dove a GPT-4 è stato sottoposto come stimolo iniziale un frammento della novella di Pirandello "La vittoria delle formiche" (questo è solo uno dei numerosi che ho sottoposto finora a GPT-4, tutti con esiti simili) il modello mostra di saper svolgere adeguatamente il compito di cooperazione interpretativa del testo; si noti, peraltro, che uno degli sviluppi proposti è quello affettivamente attualizzato nel testo. In questo esempio va detto che è probabile che la novella sia nel training set di OpenAI, sebbene GPT-4 abbia fallito tutti test di recupero puntuale della sua conoscenza sul testo specifico (ad esempio il ripristino di entità nominate mascherate in brani estratti dall'originale). Ma vorrei osservare come, ai fini della mia indagine, questo sia un problema secondario, poiché scopo dell'esperimento è testare la capacità di immaginazione narrativa del modello, e la sua conoscenza della fonte non preclude questa facoltà.

6. CONCLUSIONI

L'applicazione del concetto di "lettore modello" ai Large Language Models può essere considerata una teoria sulle capacità di questi modelli di agire come interpreti competenti di narrazioni e testi. È ormai evidente come gli LLM dispongano di una competenza linguistica estesa, acquisita attraverso l'analisi probabilistica di vasti corpora testuale. Questa competenza consente loro di leggere e "comprendere" testi (anche di considerevole lunghezza, ormai), identificando in modo generalmente accurato o almeno accettabile i temi e mostrando ottima capacità di produrre sommari e riassunti coerenti. Gli LLM sono anche capaci di elaborare narrazioni complesse, ricostruendo la fabula dall'intreccio, individuando i personaggi, e le tecniche di caratterizzazione in gioco nel testo.

Un aspetto fondamentale del lettore modello è la sua capacità di generare ipotesi interpretative. A una prima indagine gli LLM più potenti sembrano essere dotati almeno in parte di questa capacità: possono proporre diverse interpretazioni di un testo, esplorare vari potenziali sviluppi narrativi e individuare le isotopie, ovvero i temi ricorrenti e le connessioni semantiche all'interno di una narrazione. Si tratta ora di valutare in modo sistematico queste capacità, al fine di capire se e

fino a che livello i modelli linguistici generativi possano divenire strumenti utilizzabili su vasta scala nell'ambito dell'analisi letteraria basata su metodi computazionali.

BIBLIOGRAFIA

[1] Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, e Owain Evans. «The Reversal Curse: LLMs trained on 'A is B' fail to learn 'B is A.'». arXiv:2309.12288 [cs.CL], 2023. https://doi.org/10.48550/arXiv.2309.12288.

[2] Borst, Janos, Jannis Klaehn, e Manuel Burghardt. «Death of the Dictionary? – The Rise of Zero-Shot Sentiment Classification». In *Proceedings of the Computational Humanities Research Conference 2023*, 3558:303-19, 2023.

[3] Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. «Sparks of Artificial General Intelligence: Early experiments with GPT-4». arXiv:2303.12712 [cs.CL], 2023. https://doi.org/10.48550/arXiv.2303.12712.

[4] Ciotti, Fabio. «ChatGPT: un Pappagallo Stocastico può essere di aiuto a un Vero Ricercatore (Umanistico)?» In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 245-250, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[5] Ciotti, Fabio. «Minerva e il pappagallo». *Testo e Senso*, fasc. 26 (2023): 289-315.

[6] Devlin, Jacob, Ming-Wei Chang, Kenton Lee e Kristina Toutanova. «BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding». In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171-86. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-1423.

[7] D'Souza, Lyra, e David Mimno. «The Chatbot and the Canon: Poetry Memorization in LLMs». In *Proceedings of the Computational Humanities Research Conference 2023*, 3558:475-89, 2023.

[8] Eco, Umberto. *Lector in Fabula*. Milano: Bompiani, 1979.

[9] Elhage, Nelson, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, et al. «Toy Models of Superposition», 2022.

[10] Gavin, Michael. *Literary Mathematics: Quantitative Theory for Textual Studies*. California: Stanford University Press, 2023.

[11] Heuser, Ryan James. «Word vectors in the eighteenth century». In *ADHO 2017-Montréal*, 2017.

[12] Hicke, Rebecca M.M., e David Mimno. «T5 meets Tybalt: Author Attribution in Early Modern English Drama Using Large Language Models». In *Proceedings of the Computational Humanities Research Conference 2023*, 3558:274-302, 2023.

[13] Iser, Wolfgang. *L'atto della lettura: una teoria della risposta estetica.* Collezione di testi e di studi. Bologna: Il Mulino, 1987.

[14] Kaganovich, Pavel, Ophir Münz-Manor, e Elishai Ezra-Tsur. «Style Transfer of Modern Hebrew Literature Using Text Simplification and Generative Language Modeling». In *Proceedings of the Computational Humanities Research Conference 2023*, 3558:391-412, 2023.

[15] Konle, Leonard, Agnes Hilger, e Fotis Jannidis. «On Character Perception and Plot Structure of German Romance Novel». In *Proceedings of the Computational Humanities Research Conference 2023*, 3558:592-615, 2023.

[16] Kovač, Grgur, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, e Pierre-Yves Oudeyer. «Large Language Models as Superpositions of Cultural Perspectives». arXiv:2307.07870 [cs.CL], 2023. https://doi.org/10.48550/arXiv.2307.07870.

[17] Mikolov, Tomas, Chen Kai, Corrado Greg, e Dean Jeffrey. «Efficient Estimation of Word Representations in Vector Space». arXiv:1301.3781 [cs.CL], 2023. https://doi.org/10.48550/arXiv.1301.3781.

[18] Pareschi, Remo. «Abductive reasoning with the GPT-4 language model: Case studies from criminal investigation, medical practice, scientific research». *Sistemi intelligenti*, 2023, 435-44.

[19] Piantadosi, Steven T. «Modern language models refute Chomsky's approach to language». In *From fieldwork to linguistic theory: A tribute to Dan Everett. Edited by Edward Gibson and Moshe Poliak*, a cura di Edward Gibson e Moshe Poliak. Berlin: Language Science Press, 2024.

[20] Prince, Gerald. «Introduction to the Study of the Narratee Reader-Response Criticism». In *Reader-Response Criticism*, a cura di Jane P. Tompkins, 7-25. Baltimore: Johns Hopkins University Press, 1980.

[21] Rebora, Simone, Marina Lehmann, Anne Heumann, Wei Ding, e Gerhard Lauer. «Comparing ChatGPT to Human Raters and Sentiment Analysis Tools for German Children's Literature». In *Proceedings of the Computational Humanities Research Conference 2023*, 3558:333-43, 2023.

[22] Roncaglia, Gino. L'architetto e l'oracolo. Forme Digitali Del Sapere Da Wikipedia a ChatGPT. Roma-Bari: Laterza, 2023.

[23] Schmid, Wolf. «Implied Reader». In *The living handbook of narratology*, a cura di Peter Hühn. Hamburg, 2014.

[24] Underwood, Ted. «Can language models predict the next twist in a story?» *The Stone and the Shell*, 2024.

[25] Underwood, Ted. «Using GPT-4 to measure the passage of time in fiction». *The Stone and the Shell*, 2023.

I simili si attraggono. La valutazione letteraria sulle piattaforme di digital social reading

Gabriele Vezzani[1], Simone Rebora[2], Massimo Salgaro[3]
[1] Università di Verona, Italia - gabriele.vezzani@univr.it
[2] Università di Verona, Italia - simone.rebora@univr.it
[3] Università di Verona, Italia - massimo.salgaro@univr.it

ABSTRACT
La rivoluzione digitale che interessa una parte sempre crescente del campo letterario ha avuto fra i suoi effetti principali quello di attribuire al lettore comune un'inedita centralità. L'impossibilità di concepire i lettori come soggetti passivi, diretti da cause esterne, impone di superare i modelli della valutazione letteraria in cui il valore attribuibile a un'opera viene visto come il risultato della decisione di un ristretto gruppo di lettori esperti, suggerendo l'applicazione di criteri più inclusivi. In questo paper, proponiamo di considerare la valutazione letteraria come parte di un complesso sistema di relazioni tra singoli lettori e mostriamo, usando la teoria dei grafi, come le dinamiche interne a tale sistema possano portare all'emergere di criteri valutativi condivisi. Al fine di ottenere dati reali su cui basare la nostra analisi, ci si è focalizzati sulla piattaforma di digital social reading Goodreads.

PAROLE CHIAVE
Valutazione letteraria; digital social reading; teoria dei grafi; recensioni online.

1. INTRODUZIONE
Come è stato ormai notato da diversi studiosi [15, 18], nel corso degli ultimi anni il panorama letterario ha subito profonde trasformazioni dovute all'influenza delle tecnologie digitali. Per quanto riguarda la ricezione, e più in particolare la valutazione delle opere letterarie, il principale mutamento a cui si è assistito consiste nell'inedita centralità di cui gode ora il lettore comune.

Prima che essa sconfinasse nell'ecosistema digitale, la valutazione della letteratura era appannaggio quasi esclusivo di una ristretta cerchia di 'professionisti', come autori, critici letterari, ricercatori, e insegnanti. A questi soltanto era riservato il privilegio di diffondere le proprie opinioni riguardo a cosa dovesse considerarsi letteratura di qualità e cosa no, mentre i lettori 'inesperti' erano visti come una massa acritica ed omogenea, capace di esprimersi soltanto mediante l'acquisto delle opere prescritte dai critici stessi [20] o dagli ingranaggi dell'industria culturale [1]. Oggi, grazie a piattaforme come Goodreads o Wattpad, nonché alla crescente circolazione di materiale letterario su Instagram o Twitter (ora X), a questo sistema ne subentra uno estremamente più fluido, in cui ogni lettore ha la possibilità di esprimere il proprio giudizio e influenzare in maniera diretta il panorama letterario. Su queste piattaforme, l'attenzione e l'attribuzione di valore, non più dirette da organismi istituzionali, diventano le monete di uno scambio democratico ed orizzontale.

La digitalizzazione del campo letterario porta così al crollo di ogni ideale monolitico di qualità. La nuova libertà d'espressione di cui gode il lettore comune fa sì che il panorama della valutazione letteraria si popoli dei criteri più disparati, dimostrando quanto poco difendibile sia ormai l'idea di un unico canone, quale manifestazione di un ideale unico e trasversale di qualità. Appaiono ora evidenti tanto la pluralità dei canoni quanto il fatto che la validità di questi è sempre relativa ad una (o più) comunità di lettori ed è pertanto situata in un preciso contesto storico e, soprattutto, sociale. La rivoluzione digitale impone insomma di ripensare le teorie secondo cui veniva tradizionalmente compresa la valutazione delle opere letterarie, costruendo un modello capace di rendere conto del contributo di qualunque lettore indipendentemente dal suo inquadramento istituzionale, un modello che ponga al centro l'individuo, la sua interazione col libro e il contesto sociale all'interno del quale questa ha luogo.

Un simile modello dovrebbe, prima di tutto, essere in grado di dare ragione del fatto che, pur fondata sulle azioni e opinioni di singoli individui, la valutazione delle opere letterarie tenda ad organizzarsi secondo criteri condivisi in maniera più o meno omogenea da determinati gruppi di lettori. Già Bourdieu [5] notava come a diverse classi sociali corrispondessero diversi sistemi di preferenze. Seppure l'almeno apparente determinismo che nel pensiero del sociologo francese fa dipendere il gusto dalle dinamiche di classe sia stato oggetto di critica [12, 15], la presenza di criteri valutativi socialmente determinati rimane un dato difficilmente contestabile [16].

Il nostro modello dovrà quindi essere in grado di spiegare la presenza di criteri valutativi condivisi come fenomeno emergente, risultato sistemico e globale delle interazioni locali fra i singoli agenti che compongono il sistema all'interno

del quale vengono recepite le opere letterarie. Si può compiere un primo passo nell'affrontare tale problematica considerando che il nostro gusto, e pertanto anche il sistema delle nostre preferenze in ambito letterario, è fortemente legato alla nostra identità [9]. Ricerche provenienti dall'ambito della psicologia sociale [8,19] hanno mostrato come ciascuno di noi sia costantemente alla ricerca di feedback esterni in grado di confermare e legittimare i valori che, da noi interiorizzati, costituiscono la base della nostra identità. Anche il gusto può essere considerato come il risultato di un atto di interiorizzazione di valori, di natura non morale ma estetica [10]. Ciò ci aiuta a capire quanto importante sia per noi che altri condividano le nostre preferenze e idiosincrasie, così legittimando i valori che si manifestano tramite esse. Possiamo ipotizzare che la ricerca di tale legittimazione proceda, per i lettori, secondo due canali: 1) rafforzando i legami che li uniscono a soggetti che condividono i loro stessi gusti e allentando quelli con soggetti che non li condividono, contribuendo così alla formazione di comunità fondate su gusti condivisi; 2) lasciandosi influenzare, tanto nella scelta di nuovi libri da leggere che nella valutazione di libri già letti, dalle opinioni maggioritarie all'interno della comunità di appartenenza [2], contribuendo così a rafforzare l'omogeneità di quest'ultima.

Le piattaforme di *digital social reading* [17] offrono un perfetto laboratorio per testare simili teorie. In particolare, Goodreads, incorporando le funzionalità di un tradizionale *social network* e permettendo quindi ai propri utenti di stringere amicizie e seguirsi a vicenda, si presenta come il setting ideale per studiare la componente sociale della valutazione letteraria. Essendo l'intera piattaforma dedicata all'espressione dei propri gusti in fatto di letteratura, è possibile interpretare l'amicizia fra due utenti come una casistica del fenomeno generale secondo cui due individui che condividono gusti simili tendono a consolidare i legami che li uniscono. Sulla base di questo assunto, abbiamo formulato le seguenti ipotesi:

- Un network costruito seguendo i rapporti di amicizia fra gli utenti di Goodreads tenderà ad organizzarsi secondo una forte struttura modulare, rispecchiando quanto si è detto sopra riguardo l'emergere di comunità basate sulla condivisione di gusti simili (**H1**).
- All'interno di tali comunità, il gusto degli utenti sarà più omogeneo di quanto non lo sarebbe sull'intero network. In particolare, ipotizziamo che la probabilità che un utente A abbia letto gli stessi libri di un utente B sarà maggiore se A e B appartengono alla stessa comunità di quanto non lo sarebbe se fossero estratti casualmente dell'intero network (**H2a**). Inoltre, ci aspettiamo che, qualora appartengano alla stessa comunità, A e B tenderanno a valutare in maniera simile opere simili, ovvero che l'appartenenza a una data comunità abbia un effetto statisticamente significativo sulla valutazione delle opere (**H2b**).

2. METODOLOGIA

2.1. Dati

Al fine di ottenere un campione casuale di utenti, si è partiti dalla *sitemap* di Goodreads[1]. Qui, in una serie di files xml costantemente aggiornata, si trovano elencati gli utenti della piattaforma, ordinati in base all'ultimo periodo che li ha visti attivi sulla stessa. Si sono così scaricati i link alle pagine di 263,822 utenti, dai quali si è estratto in maniera casuale un campione di 1000 unità. Con uno script Python[2], servendosi del pacchetto Selenium in congiunzione con la piattaforma Dockers, sono stati salvati i link alle pagine degli utenti che figuravano come amici di ciascuno dei soggetti di partenza. La stessa operazione è stata svolta altre due volte, ogni volta partendo dal gruppo di amici creato all'iterazione precedente.

I dati ricavati mediante la procedura appena descritta sono stati organizzati nella forma di un network, nel quale ogni nodo rappresenta un utente, connesso tramite archi non-direzionali ai nodi corrispondenti alle sue amicizie. Complessivamente, il network è formato da 20,492 nodi e 23,379 archi. Per avere la possibilità di replicare eventuali risultati, abbiamo ripetuto la medesima procedura partendo da un diverso set di mille utenti (sempre estratti casualmente dagli originari 263,822, escludendo quelli selezionati nel primo round) e attivando lo script Python quattro volte invece di tre. In questo caso, si è ottenuto un network composto da 59,880 nodi e 69,750 archi.

In entrambi i casi, la distribuzione dei gradi dei nodi (il numero di archi da essi posseduti) si è rivelata fortemente asimmetrica, con una grande porzione del network (~90%) costituita da nodi aventi una sola connessione[3]. Dal momento che i legami fra lettori (qui operazionalizzati come amicizie online) rappresentano un elemento centrale della nostra

[1] https://www.goodreads.com/siteindex.user.xml

[2] Tutti gli script utilizzati per questo articolo sono disponibili al link https://github.com/GVezzani/AIUCD24.git

[3] Con ogni probabilità, ciò è dovuto prevalentemente al modo in cui si sono costruiti i grafi, scaricando, ad ogni round di scraping, le amicizie degli utenti le cui informazioni erano state scaricate nel corso del round precedente. Non potendo continuare il processo fino ad esaurire tutti gli utenti della piattaforma, necessariamente ci si è dovuti ad un certo punto arrestare. Così, non si sono scaricati i dati relativi alle amicizie degli utenti dell'ultimo round, i quali quindi figurano, nella maggior parte dei casi, come aventi un solo amico (l'utente del round precedente partendo dal quale si è arrivati a loro).

domanda di ricerca, si è deciso di concentrare ulteriori indagini soltanto su una porzione dei due grafi, escludendo da essi i nodi con troppe poche connessioni. Dopo aver tentato diverse configurazioni (vd. il grafico nella Fig. 1), è risultato che il valore ottimale per tale soglia fosse 3 per il primo network e 5 per il secondo. Per mantenere coerenza fra le due analisi e massimizzare il numero di connessioni all'interno dei network, si è scelto di escludere da entrambi i network i nodi di grado inferiore a 5. I grafi così ottenuti sono composti, rispettivamente, da 625 nodi e 1622 archi e 708 nodi e 2706 archi.

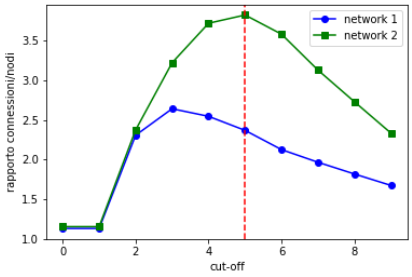


Figura 1. Numero di archi per nodo a diversi valori del cut-off.

Infine, con un secondo script Python, si sono scaricati i dati relativi allo storico delle letture di ciascun utente in entrambi i network. In particolare, sono stati salvati i titoli dei libri letti e, quando disponibile, la valutazione ad essi assegnata (da 1 a 5). In totale, gli utenti nel primo grafo hanno letto 38,538 libri diversi, quelli nel secondo 61,321.

Purtroppo, pochi di questi libri sono stati valutati da un numero di utenti tale da rendere possibile (o comunque significativa) un'analisi statistica. Per fare un esempio, un modello di regressione a effetti misti (si veda sotto per maggiori dettagli) richiede che si abbiano almeno cinque osservazioni per ogni livello di ciascuno degli effetti random inclusi. I libri che soddisfano tale condizione, e che quindi avrebbero potuto essere analizzati con tale metodo, non sono che il 3% del totale. Per ovviare a questa impasse, si è deciso di aggregare le opere per genere letterario.

Le informazioni relative al genere di un'opera sono presenti su Goodreads sotto forma di tags generati dagli utenti stessi. Nel corso dello scraping di questi dati si è incorsi in un problema affine a quello appena esposto: per la maggior parte delle opere nel nostro corpus nessun utente aveva generato alcun tag. Ciò ha reso per noi impossibile determinare il genere di 84% dei libri nel nostro corpus. Come è evidente dalla statistica appena citata, nonostante le sue limitazioni, questo approccio ci avrebbe permesso di lavorare con più opere di quante ne avremmo avute a disposizione basando le nostre analisi sui singoli libri, e si è scelto pertanto di aggregarle per genere in fase di analisi, escludendo quelle prive di genere.

2.2. Analisi

Al fine di testare l'ipotesi **H1**, entrambi i grafi sono stati analizzati mediante l'algoritmo di Louvain [4]. In breve, questo metodo è usato per trovare le comunità in cui si suddivide un network, ovvero gruppi di nodi all'interno dei quali si ha una densità di connessioni maggiore rispetto a quella che si otterrebbe selezionando casualmente un insieme di nodi della stessa numerosità. Il coefficiente di modularità, che per un grafo non pesato e non direzionale come i nostri può variare da un minimo di -0.5 a un massimo di 1 [7], misura la forza della struttura modulare del grafo.

Per testare l'ipotesi **H2a**, per ciascun utente di ciascuno dei due network si è poi calcolato:

- La probabilità di aver letto almeno un libro in comune con un altro utente all'interno della stessa comunità (gruppo 1) e la probabilità di aver letto almeno un libro in comune con un altro utente casuale (gruppo 2).
- Il numero medio di letture in comune con utenti appartenenti alla stessa comunità (gruppo 1) e il numero medio di letture in comune con utenti dell'intero network (gruppo 2).

L'influenza che l'appartenenza a una data comunità esercita sulle valutazioni delle opere letterarie (**H2b**) è stata verificata sia globalmente, sull'interezza di entrambi i grafi, che concentrandosi su esempi rappresentati da specifici generi letterari. In primo luogo, si è creato un dataset delle valutazioni, elencando per ciascuna l'identificativo dell'utente e della comunità d'appartenenza, il genere letterario del libro valutato e la valutazione stessa (da 1 a 5). Servendosi del pacchetto lme4 di R [3], si è costruito un modello di regressione lineare a effetti misti, al quale si è richiesto di predire le valutazioni degli utenti basandosi su un'intercetta e sugli effetti random rappresentati da: i singoli lettori, il genere dei libri valutati e, per gli stessi generi, le comunità di appartenenza degli utenti (effetto random innestato nel precedente). In parole povere, l'intuizione alla base del modello è che: a) diversi lettori valutano secondo criteri diversi, b) generi diversi vengono valutati diversamente e c) utenti delle stesse comunità tendono a valutare in modo simile libri appartenenti a un medesimo genere letterario. Naturalmente, il terzo punto è dove risiede il nostro interesse. Per valutarne la significatività statistica si è usata la funzione 'anova' di R, confrontando le performance del nostro modello con quelle di un secondo, identico al precedente se non per l'esclusione dell'effetto di nostro interesse. L'intento, qui, consisteva nel verificare se l'aggiunta della variabile relativa alle comunità di appartenenza degli utenti comportasse un miglioramento significativo rispetto alla performance un modello che includesse come effetti random soltanto quelli relativi agli utenti e ai generi letterari dei libri valutati.

Infine, per ciascuno dei generi letterari nel dataset è stato eseguito un test ANOVA per verificare se ci fossero delle differenze statisticamente significative fra le valutazioni di utenti appartenenti a diverse comunità.

3. RISULTATI

Per quanto riguarda l'ipotesi **H1**, l'impiego dell'algoritmo di Louvain ha permesso di dividere con successo entrambi i grafi in comunità di utenti. Nel primo grafo, sono state individuate 70 comunità, per un coefficiente di modularità pari a 0.8. Nel secondo, si sono trovate 28 comunità, con un coefficiente di 0.61.

Figura 2. Porzione del secondo grafo con comunità. La visualizzazione è stata realizzata col software Gephi, usando l'algoritmo ForceAtlas 2

Simili risultati sono stati ottenuti anche per l'ipotesi **H2a**. Il primo test di Mann-Whitney sul primo grafo ha dato risultati statisticamente significativi (u = 273,279, $p < 0.001$, $r_{pbi} = 0.3$), dimostrando che la probabilità di aver letto almeno un libro in comune è maggiore di 0.11 per utenti appartenenti alla stessa comunità di quanto non lo sia per utenti casuali. Il secondo test di Mann-Whitney ha dimostrato (u = 262,937, $p < 0.001$, $r_{pbi} = 0.14$) che due utenti appartenenti allo stesso gruppo hanno letto, in media, 0.5 libri in comune in più rispetto a due utenti casuali. Come si può notare nella Tabella 1, tali risultati sono stati confermati anche per il secondo grafo.

In riferimento all'ipotesi **H2b**, il modello lineare a effetti misti è stato adattato nel primo grafo utilizzando *Restricted Maximum Likelihood estimation* (REML) e il criterio di convergenza è stato raggiunto con un valore di 17,306. Per quanto riguarda gli effetti di nostro interesse, la varianza nelle valutazioni dovuta al genere del libro valutato equivale a 0.009 (std. dev. = 0.09), mentre quella relativa alla comunità ('genere:comunità', in quanto siamo interessati ai voti di utenti appartenenti a comunità diverse su libri dello stesso genere) a 0.02 (st. dev. = 0.14). Il confronto con un modello dal quale era stato escluso il secondo degli effetti menzionati, ha dimostrato che la sua aggiunta comporta un miglioramento di performance statisticamente significativo ($X^2 = 24.03$, $p < 0.01$).

Per alcuni generi, il test ANOVA ha dato risultati statisticamente significativi. È il caso, per fare soltanto due esempi, dei generi 'Historical Fiction' (F (45, 288) = 1.7, $p < 0.01$) e 'Romance' (F (41, 1092) = 4.0, $p < 0.01$). Tuttavia, per ragioni che verranno discusse in seguito, nel 74% dei casi è stato impossibile raggiungere la soglia di significatività.

Anche in questo caso (cfr. Tabella 1) la replicazione sul secondo grafo ha confermato i risultati ottenuti.

Test	Risultati
Mann-Whitney 1	u = 333,747, $p < 0.001$, $r_{pbi} = 0.38$, differenza fra medie = 0.13
Mann-Whitney 2	u = 293,454, $p < 0.001$, $r_{pbi} = 0.07$, differenza fra medie = 0.35
Modello a effetti misti	'genere: varianza = 0.02(st. dev. = 0.15), 'genere:comunità': varianza = 0.005 (st. dev. = 0.07)
Confronto fra modelli	$X^2 = 4.72$, p = 0.02
ANOVA	Risultati statisticamente significativi nel 24% dei casi

Tabella 1. Risultati dei test eseguiti sul secondo network.

4. DISCUSSIONE

I risultati dell'applicazione dell'algoritmo di Louvain a entrambi i network esaminati confermano l'ipotesi **H1**, mostrando come la rete di interazioni fra gli utenti di Goodreads si organizzi per strutture modulari, o per comunità. È bene notare che, dal momento che Goodreads non fornisce dettagliate informazioni demografiche relative ai propri utenti, non è possibile escludere categoricamente l'ipotesi che i gruppi trovati siano il riflesso di fattori esterni, come ad esempio il livello di istruzione o la classe sociale dei soggetti. Proprio perché tali informazioni non sono accessibili sulla piattaforma, tuttavia, possiamo presumere che esse non svolgano un ruolo preponderante nel determinare i comportamenti degli utenti (fra cui anche la scelta di stringere amicizie), e che questi siano basati prevalentemente sui dati che effettivamente circolano sul sito, ovvero dati relativi alla valutazione di opere letterarie. Tale ipotesi è in linea col secondo dei risultati della nostra immagine, ovvero l'aver mostrato come la struttura del network di relazioni sociali su Goodreads rispecchi differenze di gusti e criteri valutativi.

Ciò è evidente, in primo luogo, nello storico delle letture degli utenti. Come giustamente notano Heydebrand e Winko [13], la scelta di un libro da leggere può essere considerata come un atto valutativo a tutti gli effetti. A questo proposito, i risultati di entrambi i test di Mann-Whitney dimostrano, confermando l'ipotesi **H2a**, che le scelte di due utenti appartenenti alla stessa comunità sono più coerenti di quanto non lo siano quelle di due utenti casuali. In secondo luogo, i risultati del modello di regressione lineare a effetti misti dimostrano, confermando l'ipotesi **H2b**, che gli individui appartenenti alla stessa comunità tendono a condividere simili criteri valutativi, e quindi a valutare opere appartenenti allo stesso genere in maniera più coerente di quanto non lo farebbero due utenti casuali.

Tuttavia, i risultati dei test ANOVA impongono cautela nell'interpretazione. Il fatto che sia stato possibile trovare risultati statisticamente significativi soltanto per una porzione dei generi presi in considerazione riflette, infatti, la complessità dell'effetto in esame. È innanzitutto possibile che, nel caso di molti generi, semplicemente non si sia raggiunto un numero di lettori per comunità abbastanza elevato da garantire la significatività statistica dei risultati. Due potenziali fonti di variabilità devono essere considerate. Per primi, i generi stessi. Per fare soltanto un esempio, una comunità di utenti che condividono la passione per la science fiction sarà molto più omogenea nelle sue valutazioni di generi affini (distopia, fantasy…) e lascerà ai suoi membri molta più 'libertà' nel valutare opere di generi più distanti (poliziesco, romanzo storico…). La stessa variabilità si può ipotizzare, poi, in relazione agli utenti. Bisogna considerare infatti che, nella realtà, gli individui possiedono gusti poliedrici e non possono essere pensati come appartenenti a un singolo gruppo di lettori. Questo fa sì che, nelle comunità da noi individuate, siano presenti membri più o meno 'eretici', più o meno in linea con gli standard del gruppo.

A seguito di simili considerazioni, potremmo aspettarci che ogni genere, raggiunta una numerosità campionaria tale da controllare per i fattori di variabilità appena esposti, presenti differenze significative nelle valutazioni attribuitegli da lettori appartenenti a diverse comunità. Ciò, tuttavia, non avviene. Nella Figura 3, vediamo a confronto i diagrammi a scatola delle valutazioni relative ai generi 'LGBT' e 'Spirituality'. Gli utenti che hanno valutato libri afferenti al secondo sono un numero maggiore (19) di quelli che hanno fatto lo stesso per libri afferenti al primo (16). Tuttavia, il genere 'LGBT' permette di rilevare una differenza statisticamente significativa fra le valutazioni provenienti da diverse comunità di utenti, riferita in particolare alla comunità numero 36, che sembra non apprezzare particolarmente questi libri. Al contrario, per quanto riguarda il genere 'Spirituality', tutte le valutazioni si assestano fra il 4 e il 5, dimostrando una maggiore concordia. È chiaro, insomma, che l'ambiguità dei risultati ottenuti non dipende solamente da questioni legate alla numerosità campionaria. Piuttosto, sembra che essa sia dovuta alle comunità stesse e alle caratteristiche dei generi valutati. Alcuni di questi, infatti, si dimostrano particolarmente adatti a catturare le differenze di gusto fra determinate comunità, mentre altri non svolgono alcun ruolo significativo in tal senso.

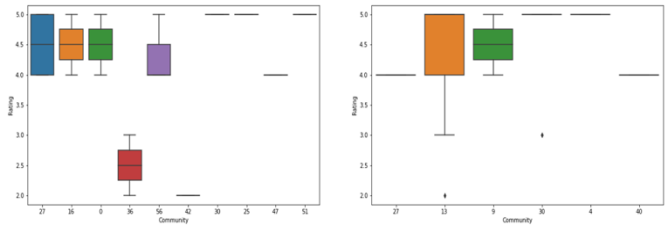


Figura 3. Confronto fra il diagramma a scatola relativo alle valutazioni per i generi 'LGBT' (a sinistra) e 'Spirituality' (a destra)

Come interpretare un simile risultato? Vicini, in questo, alla lezione di Bourdieu [6], possiamo considerare il panorama artistico come un campo. In ogni sua fase, tale campo può essere immaginato come uno spazio le cui dimensioni corrispondono a determinate opere, determinati generi o determinati movimenti letterari (in base al livello di risoluzione che si sceglie per la propria analisi). Un certo gusto altro non sarebbe che una *posizione* occupata in tale spazio, o, meglio, una presa di posizione rispetto alle dimensioni (opere, generi, movimenti) sulle quali esso è definito. La multidimensionalità del campo letterario è ciò che permette che due gusti siano in parte sovrapposti, pur rimanendo distinti: nulla, infatti, impedisce a due persone (o a due gruppi) di avere simili pareri relativamente a un certo genere, ma differire fortemente nel valore che attribuiscono a un altro.

I nostri risultati dimostrano come la presenza di criteri valutativi condivisi non debba necessariamente (nonostante, talvolta, possa) essere ricondotta all'azione di agenti istituzionali o determinanti socioeconomiche. Essa può anche essere interpretata come il risultato della libera azione di singoli individui che, esplorando il campo letterario e confrontandosi con altri individui impegnati nella medesima attività, contribuiscono all'emergere di una omologia fra le posizioni da loro occupate nel campo stesso (gusti) e nella rete di rapporti sociali che li unisce gli uni agli altri.

I processi alla base di una tale omologia possono essere di due tipologie principali:

1) processi di selezione, in base ai quali gli individui tendono a stringere legami sociali con coloro che possiedono gusti simili ai loro (pensiamo al caso di un utente che manda una richiesta d'amicizia all'autore di una recensione appena letta e con la quale si trova particolarmente d'accordo);
2) processi di influenza, in base ai quali gli individui tendono ad omologare i propri gusti a quelli dei membri dei gruppi sociali a cui appartengono (banalmente, ciascun utente sarà maggiormente disposto ad accettare i suggerimenti o le suggestioni provenienti dalla sua cerchia di amici, piuttosto che da sconosciuti).

Diversi studiosi, nell'ambito della sociologia [11, 14], hanno cominciato ad esplorare come queste due tipologie di processi intervengano nella determinazione del gusto estetico. Tuttavia, tali ricerche ignorano il caso della letteratura, concentrandosi piuttosto su musica e televisione. Inoltre, manca ancora un interesse rivolto specificamente a forme di socialità digitale, come quelle considerate nel presente studio. Un possibile sviluppo futuro per questo lavoro potrebbe

consistere proprio nel determinare in che misura i fenomeni qui individuati debbano essere attribuiti a dinamiche di selezione e quanto, invece, a dinamiche di influenza.

BIBLIOGRAFIA

[1] Adorno, Theodore, e Max Horkeimer. *Dialectic of enlightenment*. London: Blackwell Verso, 1997.

[2] Asch, Solomon. «Effects of group pressure upon the modification and distortion of judgments." Groups, Leadership and Men». In *Human Relations*, 177-190. Lancaster: Carnegie Press, 1971.

[3] Bates, Douglas, Martin Mächler, Bolker Ben, e Steve Walker. «Fitting Linear Mixed-Effects Models Using lme4». *J. Stat. Soft.* 67 (2015): 1-48.

[4] Blondel, Vincent, Guillaume Jean-Loup, Renauld Lambiotte, e Etienne Lefebvre. «Fast unfolding of communities in large networks». *J. Stat. Mech.* 2008, 1-12.

[5] Bourdieu, Pierre. *La Distinction: Critique sociale du jugement*. Paris: Minuit, 1979.

[6] Bourdieu, Pierre. *Les règles de l'art: genèse et structure du champ littéraire, Nouv. éd., revue et corrigée*. Paris: Éd. du Seuil, 2010.

[7] Brandes, Ulrik, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, e Dorothe Wagner. «On Modularity Clustering». *IEEE Trans. Knowl* 20 (2008): 172-188.

[8] Burke, Peter, e Donald Reitzes. «An Identity Theory Approach to Commitment». *Social Psychology Quarterly* 54 (1991): 239-251.

[9] Dolby, Nadine. «The Shifting Ground of Race: The role of taste in youth's production of identities». *Race Ethnicity and Education* 3 (2000): 7-23.

[10] Fingerhut, Joerg, Javier Gomez-Lavin, Claudia Winklmayr, e Jesse Prinz. «The Aesthetic Self. The Importance of Aesthetic Taste in Music and Art for Our Perceived Identity». *Front. Psychol* 11 (2021).

[11] Friemel, Thomas. «Network dynamics of television use in school classes». *Social Networks* 34 (2012): 346-358.

[12] Hennion, Antoine. «Those Things That Hold Us Together: Taste and Sociology». *Cultural Sociology* 1 (2007): 97-114.

[13] Von Heydebrand, Renate, e Simone Winko. *Einführung in die Wertung von Literatur: Systematik - Geschichte - Legitimation*. München: Schöningh, 1996.

[14] Lizardo, Oscar. «How Cultural Tastes Shape Personal Networks». *Am Sociol Rev* 71 (2006): 778-807.

[15] Murray, Simone. *The digital literary sphere: reading, writing, and selling books in the Internet era*. Baltimore: Johns Hopkins University Press, 2018.

[16] Prior, Nick. «Critique and Renewal in the Sociology of Music: Bourdieu and Beyond». *Cultural Sociology* 5 (2011): 121-138.

[17] Rebora, Simone, Peter Boot, Federico Pianzola, Brigitte Gasser, J.Berenike Herrmann, Maria Kraxenberger, Moniek M. Kuijpers, et al. «Digital humanities and digital social reading». *Digital Scholarship in the Humanities* 36 (2021): ii230-ii250.

[18] Salgaro, Massimo. «Literary value in the era of big data. Operationalizing critical distance in professional and non-professional reviews». *Journal of Cultural Analytics* 7 (2022).

[19] Stets, Jan, Scott Savage, Peter Burke, e Phoenicia Fares. «Cognitive and Behavioral Responses to the Identity Verification Process». In *Identity and Symbolic Interaction*, (a cura di) Richard Serpe, Robin Stryker, e Brian Powell, 65-88. Springer International Publishing, 2020.

[20] Van Rees, Kees. «How a literacy work becomes a masterpiece: On the threefold selection practised by literary criticism». *Poetics* 12 (1983): 397-417.

Il *Distant reading* è l'ornitorinco

Pietro Mazzarisi
Università degli studi di Trieste, Italia - pietro.mazzarisi@gmail.com

ABSTRACT

Negli studi di *Distant reading*, l'oggetto base su cui vengono condotte le ricerche è composto dai testi letterari e gli aspetti, caratteristiche, costruzioni e categorie linguistiche al loro interno. L'invalso che si è sviluppato è di non operare alcuna distinzione all'interno dei *corpora* su cui vengono condotte le indagini. Ai fini di validità statistica l'invalso è corretto, pena sarebbe, diversamente, la non rappresentabilità del campione. Ma ai fini dell'attività critico-interpretativa l'invalso non riconosce almeno due aspetti fondamentali: (i) tutte le classificazioni di linguistica testuale pongono il testo letterario a sé, riconoscendo a un tale oggetto base di ricerca caratteristiche peculiari e strutturali; (ii) aspetti, costruzioni, categorie e caratteristiche linguistiche oggetto base di indagine acquisiscono salienze diverse quando si presentano nella diegesi e quando si presentano nella mimesi di una stessa opera, poiché le porzioni testuali diegetiche e le porzioni testuali mimetiche di uno stesso testo letterario hanno salienze diverse. L'intervento propone una riflessione e proposta teorico-metodologica per ovviare a tali criticità, seguita da una sua applicazione concreta su un *corpus* di letteratura italiana premiata.

PAROLE CHIAVE

Distant reading; linguistica testuale; diegesi/mimesi; letteratura premiata.

1. INTRODUZIONE

Una delle monografie più recenti sul *Distant reading* lo rinomina *Literary mathematics* e lamenta la scarsa attenzione teorica che la *Cultural analytics*, le *Digital humanities* e questo specifico sottocampo hanno indirizzato verso l'oggetto base d'indagine, richiamando l'attenzione di ricercatori e ricercatrici sullo studio del *corpus* in sé: «what literary scholars bring to the interdisciplinary table is their robust and sustained attention to textual forms. We have elaborate theories for poetic and novelistic structures, yet cultural analytics and the digital humanities have proceeded to date without a working theory of the corpus as an object of inquiry» [9: 6]. Dopo aver riaffermato le persistenti necessità e mancanze di quadri teorici per il *Distant reading* [5], lo studio individua in questi termini le domande di ricerca atte a colmare queste mancanze: «We need a theory of quantitative literary analysis. Such a discourse would ask: What kinds of things are the objects of textual computation? […] A much better question to begin with is this one: What relations exist between the corpus and its source texts?» [9: 165-166]. Nel rispondere a queste domande di ricerca, lo studio di Gavin ha il pregio di portare la necessaria e dovuta attenzione verso l'intersecazione, all'interno delle pratiche di *Distant reading*, tra le analisi computazionali dei testi letterari e dei metadati; questi ultimi intesi nell'accezione più ampia che superi le sole autorialità, anno, luogo e casa editrice di pubblicazione e si soffermi anche sulle vicende inter/intrapersonali di tali autorialità non dimenticando le eventuali e ulteriori differenze etniche, di genere e di altra natura. In questa direzione, lo studio ricostruisce la nascita dell'ipotesi distribuzionale di Zellig Harris [12] dagli studi linguistici firthiani [8] per estenderla alle pratiche di *Distant reading*, dimostrando infine l'assunto secondo il quale «*Similar words tend to appear in documents with similar metadata* […]. Similar words will be used by similar authors, at similar times, when describing similar things in similar places» [9: 175].

Come la maggior parte degli studi di *Distant reading*, anche il sopraccitato studio manca però di un passaggio preliminare e fondamentale: ovvero la necessità di fare un passo indietro e non trattare il testo letterario alla stregua di un qualsiasi altro testo. Da decenni, la linguistica testuale ha infatti proposto classificazioni ampiamente note e condivise e che, pur basandosi su differenti prospettive tipologiche, tutte distinguono il testo letterario da altre forme testuali: nelle prospettive funzionali più note è definito "testo narrativo" [22] e "testo letterario e poetico" [6], "testo a libertà interpretativa alta" [19] nella prospettiva interpretativa.

Il testo letterario è un mammifero, condivide caratteristiche comuni a tutte le altre tipologie testuali. Il testo letterario è un ornitorinco, come le echidne e a differenze di tutti gli altri mammiferi possiede caratteristiche peculiari. Ipotizzando uno studio sulla riproduzione nei mammiferi, si dovrebbero trattare diversamente i monotremi da tutti gli altri mammiferi sulla base della oviparità dei primi rispetto alla viviparità dei restanti. A discapito di tali peculiarità, fino ad oggi l'invalso nelle ricerche di *Distant reading* è stato studiare il testo letterario come un testo comune. Malgrado la specificità sottolineata dalla linguistica testuale e la forte attenzione alle forme testuali della ricerca letteraria rilevata nello studio di Gavin, nelle proprie pratiche il *Distant reading* non ha riconosciuto che i *corpora* su cui concentra gli studi sono composti di testi strutturalmente diversi dai testi dei *corpora* impiegati negli studi politici, giuridici, psicologici ecc.

Gli studi più entusiasti che hanno seguito la nascita del concetto di *Distant reading* sviluppandolo e mettendolo in pratica si sono spesso proposti come approcci coincidenti con la condizione descritta da Umberto Eco a cui echeggia il titolo di questo intervento [7]: scoperte che metterebbero in discussione categorizzazioni ritenute ormai storicamente organiche. Ma la maggior parte di essi sembra essersi comportata come l'ornitorinco, almeno secondo una visione antropocentrica sulla coscienza: ignaro della propria specificità rispetto a tutti gli altri mammiferi. Un esempio recente può essere lo stesso studio sopraccitato: oltre al proporre la definizione di *Literary mathematics* in sostituzione di *Distant reading* e a non concentrare la necessaria attenzione sulle specificità del testo letterario, ha svolto le proprie indagini su un *corpus* costituito non solo di testi letterari, ma anche da testi di critica letteraria in aggiunta ai testi letterari [9].

2. DIEGESI E MIMESI

Nella visione del *Distant reading* qui proposta, la disciplina dovrebbe anzitutto riconoscere e rispettare l'oggetto base della ricerca e le sue peculiarità, ovvero le peculiarità del testo letterario. Di conseguenza, la creazione di un quadro teorico base dovrebbe porsi come principale domanda la seguente: come si studia il testo letterario computazionalmente entro un quadro teorico che (i) alla fonte ne riconosca la specificità testuale e (ii) sia anche operazionalizzabile [14]?

Per rispondere a questa domanda, l'approccio proposto propone di ripartire dalla millenaria opposizione tra diegesi (διήγησις) e mimesi (μίμησις). La tradizione occidentale fa risalire a Platone le prime categorizzazioni e distinzioni tra ciò che è riconducibile al narratore e ciò che, invece, è riconducibile ai personaggi. Ne *La Repubblica*, Platone segue la concezione greca per cui l'arte e la poesia sono forme di imitazione, ma ne offre un approfondimento attraverso il confronto con Adimanto quando gli chiede se, in definitiva, tutti i racconti fatti da mitologi e poeti non si riducano a una narrazione diretta, imitativa o delle due forme intrecciate [18: III, 392d]. Per esemplificare la distinzione, Platone parafrasa lo scambio dialogico tra il sacerdote apollineo Crise e i carcerieri della figlia all'inizio dell'*Iliade*, offrendo così una riformulazione degli scambi dialogici in diegesi e sottolineando la distinzione tra discorso diretto e discorso indiretto [18: III, 393d-394b]. Infine, àncora questa distinzione principalmente a quattro generi letterari: «Nell'invenzione poetica c'è un genere completamente imitativo, come tu dici, ed è rappresentato dalla tragedia e dalla commedia. Poi ce n'è un altro in cui è il poeta stesso a narrare, come accade particolarmente nei ditirambi. Infine c'è un terzo tipo, misto di narrazione e di imitazione, che si trova nella poesia epica e in molti altri componimenti. Mi capisci, non è vero?» [18: III, 394c].

La riflessione teorica sull'opposizione diegesi/mimesi è millenaria, nella tradizione occidentale essa parte già da Aristotele [1: I-II, 1447a-1448a] e continua a svilupparsi fino alla narratologia classica [10: 52-67, 11: 71-72] e oltre. Infatti, per l'attività critico-interpretativa letteraria risulta fondamentale poter continuare a contare sulle categorie della diegesi e della mimesi, poiché su queste ha costruito una buona fetta dei suoi strumenti critici, dallo *skaz* di Ejchenbaum, per l'analessi di Genette alla polifonia di Bachtin, solo per citare pochi esempi. Il *Distant reading*, invece, non ha ereditato l'opposizione diegesi/mimesi, privando la disciplina di una basilare riflessione teorica, benché risulti fondamentale per l'attività critico-interpretativa non solo, per esempio, ottenere un *dataset* con il tipo di lessico che caratterizza un *corpus* letterario, ma poter anche distinguere quale delle due dimensioni tra diegesi e mimesi è caratterizzata da quale tipo di lessico e incrociare, raffrontare, interpretare il differenziale tra i dati.

3. APPROCCIO TEORICO-METODOLOGICO: ASPETTI E LIMITAZIONI

Il nuovo approccio teorico-metodologico, denominato "analisi dell'asse diegetico-mimetico" (*Diegetic-Mimetic Axis Analysis*), ha per obiettivi (i) riconoscere la specificità del testo letterario operazionalizzando alla sorgente l'opposizione teorica diegesi/mimesi, (ii) preservare la validità statistica dei campioni di indagine e (iii) creare le condizioni per poter ottenere una miglioria nella raccolta dati, arricchendoli qualitativamente in granularità e salienza. Per raggiungere questi obiettivi la procedura prevede una previa partizione automatica dei testi letterari in porzioni testuali diegetiche e in porzioni testuali mimetiche con tecnica di *text mining*. La partizione automatizzata esclude l'intervento e la discrezione umana, basandosi su elementi interni dei testi e predeterminati da chi ha creato i testi: i segni di interpunzione e la convenzioni formali a cui la maggior parte dei testi letterari in prosa ricorre per segnare l'avvicendamento tra diegesi e mimesi. Pertanto, la validità statistica del campione non viene compromessa, poiché esso viene segmentato e non modificato.

La procedura prevede poi la conduzione di analisi computazionali sulle distinte porzioni testuali così ottenute. Infine, prevede la conduzione di analisi differenziali tra le distinte entità testuali ottenute, ovvero tra quella diegetica e quella mimetica. L'analisi differenziale è un tipo di indagine preso in prestito da altri campi del sapere con la finalità di migliorare gli aspetti interpretativi di fronte a un problema. Nei campi della diagnostica medica e psicologica include la presa in considerazione di fenomeni diversi per giungere a una determinazione quanto più corretta possibile. In economia compara analisi di voci di bilancio diverse per determinare la sostenibilità di una scelta. Nella critica letteraria computazionale mira a resocontare come e in che misura aspetti, caratteristiche, costruzioni e categorie linguistiche oggetto base di ricerca si

manifestino nelle due dimensioni della diegesi e della mimesi e quali aspetti è possibile cogliere dalle diverse manifestazioni. In sintesi, l'analisi differenziale è un procedimento utile nel formare addizionali interpretativi per l'espressione dei giudizi critici.

Il binomio teorico diegesi/mimesi è impiegato con accezioni operative in prospettiva di operazionalizzazione e connota la diegesi come "la porzione testuale riconducibile a chi narra", la mimesi come "la porzione testuale riconducibile ai personaggi". L'approccio ingloba le riflessioni teoriche che si sono sviluppate in narratologia da tale dicotomia [10: 52-67], ma semplifica le distinzioni teoriche alle due sole dimensioni diegetica e mimetica per le necessità pratiche che si presentano quando si deve tradurre una teoria letteraria in una sequenza di istruzioni formali da comunicare con il linguaggio di programmazione; ovvero quando si deve operazionalizzare una teoria letteraria affinché le misurazioni sui testi avvengano alla luce di quel concetto letterario [14].

Di conseguenza, la metodologia ha delle limitazioni circa la propria applicabilità. La prima e più ovvia limitazione riguarda il fatto che l'approccio delimita da sé il proprio campo di applicazione, perimetrandolo alla testualità letteraria in prosa, così come la linguistica testuale suddivide il testo letterario in prosa e poesia [6]. La seconda limitazione riguarda il fatto che se la maggior parte dei testi letterari sono suddivisi in mimesi/diegesi e di norma queste due dimensioni sono graficamente indicate con segni interpuntivi, esistono eccezioni che a volte costringono un intervento umano quantomeno di controllo su come effettivamente è stato reso l'avvicendamento tra mimesi e diegesi (nel *corpus* preso in analisi nel seguente paragrafo un caso su centotré). Per ovviare a tali problematicità, la metodologia nasce come un approccio di natura *mixed method*. In questa prospettiva (e)segue il concetto di *scalable reading* proposto da Martin Mueller - che consiste in una alternanza di fasi di letture ravvicinate a fasi di letture a distanza [15]. La metodologia va così a coincidere con il fine indicato da Andrew Piper sul versante di un ricomponimento con la critica classica. Piper, infatti, ha proposto il carattere circolare e dunque ermeneutico dei passaggi da micro-analisi a macro-analisi e di nuovo a micro come base di un *mixed method* che consiste in una ermeneutica computazionale fatta di letture a fasi alterne tra vicino e lontano e altrettante alterne analisi tra i dati qualitativi e quantitativi [17]. Un approccio *mixed method* oltre a risanare la frattura con la critica classica può, inoltre, far progredire la disciplina del *Distant reading*, passando appunto dai punti intermedi e inevitabili quando una disciplina giovane si trova nelle sue fasi iniziali, dimostrando che se dapprima non si è in grado di trattare con efficienza teorica le scale delle centinaia e delle migliaia, difficilmente si potranno trattare teoricamente ed efficacemente le scale di milioni di testi letterari (e, in ogni caso, l'esperienza maturata nelle scale inferiori giova alle superiori).

4. APPLICAZIONE SU *CORPUS* DI LETTERATURA PREMIATA

L'approccio viene applicato su un *corpus* di letteratura italiana premiata composto da tutte le opere narrative scritte in lingua italiana vincitrici dei premi Strega, Campiello e Bancarella dal 1980 al 2020 per mostrare come possa migliorare qualitativamente la raccolta e interpretazione dei dati aumentandone le caratteristiche in granularità e salienza. L'intervallo storico inserisce il *corpus* nell'epoca del postmoderno. I premi letterari dovrebbero, ipoteticamente, intercettare parte delle opere che in futuro verrà inclusa nel canone letterario; questo a sua volta dovrebbe riflettere il periodo storico-letterario in cui tali opere sono state concepite. Ma in che modo, effettivamente, i premi letterari si pongono in relazione agli stili che riflettono il postmoderno? Una caratteristica del postmoderno sarebbe l'impiego di un linguaggio alleggerito da norme e restrizioni, non controllato da una mente unificatrice tramite un processo di centralizzazione, ma che rifletta una maggiore spontaneità. I premi hanno apprezzato lo stile postmoderno alleggerito da norme e restrizioni o si sono volti indirizzando la eco da loro prodotta verso gli stili "alti, controllati, sublimi" dove, proporzionalmente, è stata vigilata l'insorgenza di attenuazioni da norme e restrizioni?

Per rispondere alle domande di ricerca e mostrare la miglioria apportata dall'approccio, nel *corpus* viene preso in analisi l'impiego del *presente pro-futuro*. Una marcatezza che nella mimesi rifletterebbe la scelta di rendere un linguaggio più autentico e spontaneo, nella diegesi uno stile narrativo alleggerito da norme e restrizioni. Infatti, per Berruto, ormai «*verrò domani* risulta quasi funzionare da forma enfatica rispetto a *vengo domani*» [2: 70]. Ricercare su testi letterari scritti in lingua italiana ha portato a escludere dal *corpus* alcune opere vincitrici del premio Bancarella: alcune perché scritte in lingua straniera, dunque traduzioni non esattamente cumulabili con una ricerca sull'uso di una forma marcata; altre perché di *non-fiction*. Strega e Campiello hanno sempre e solo premiato opere di autori e autrici italiane e nell'intervallo storico non figurano testi poetici tra i vincitori. Nel 2005, il Campiello registra una vincita *ex aequo* tra Pino Roveredo (*Mandami a dire e altri racconti*) e Antonio Scurati (*Il sopravvissuto*). È stato preferito Roveredo secondo il criterio di maggiore inclusione e diversità all'interno del *corpus*. Infatti, Scurati è già presente con un'altra opera (*M. Il figlio del secolo*, Strega 2019); Roveredo no e, conseguentemente, sarebbe stato del tutto escluso. La raccolta di racconti di Roveredo è stata misurata come libro, non suddividendola in sotto-testi e lo stesso criterio è stato applicato alle altre raccolte di racconti vincitrici (ovvero 1982S, 1997S, 1999S). In questo modo, la dimensione finale del *corpus* è di 103 opere e 1.762.929

occorrenze totali di forme verbali (41 opere e 837.639 occorrenze Strega, 41/606.712 Campiello, 21/318.578 Bancarella; lo Strega include *La scuola cattolica* di Edoardo Albinati che da solo apporta 149.228 occorrenze. Vd. Fig. 1?).
Passando al presente in funzione futura, sembra essere diffuso e trasversale nelle lingue naturali, indipendentemente dalla disponibilità di tempi preposti all'espressione della futurità [3: 97-98]. Attestato già nell'italiano antico [13: 75], in epoca contemporanea è frequentemente coadiuvato da avverbi ed espressioni di tempo [16: 99] e solitamente circoscritto alla lingua colloquiale [20: 467]. Per escludere occorrenze di presente con funzioni indesiderate - come, per esempio quella storica o iterativa - il *data cleaning* ha limitato l'output alle *key words in context* (KWIC) circoscrivendolo alle co-occorrenze di presente semplice in concomitanza degli avverbi di tempo *oggi*, *stamattina*, *stasera*, *stanotte*, *domani*, *indomani*, *domattina*, *dopodomani* nelle precedenti e seguenti 17 parole.
Passando ai premi letterari, un recente studio di Simonetti prende in analisi soprattutto vittorie e cinquine finaliste degli ultimi sei anni, notando punti comuni nelle vittorie dello Strega. Sarebbero opere inclusive che non impegnano eccessivamente chi legge e propongono sempre la conformità alle coeve idee politiche. Si porrebbero dunque lontano dallo sperimentalismo, scevre di fratture allo standard, sperimentando al massimo nell'ibridazione dei generi. Sempre a supporto dei deboli, tradirebbero, una falsa coscienza poiché questa *pietas* non si traduce in solidarietà stilistica e autentica adesione linguistica al mondo popolare cui esprimono compassione [21: 163]: «La lingua di questo romanzo non deve essere troppo complicata o resistente alla decifrazione, soprattutto nella struttura sintattica […]. Lo sperimentalismo formale nel complesso continua a essere bandito, come pure […] gli stereotipi della narrativa di consumo; mentre giocare a mescolare generi diversi […] viene senz'altro incoraggiato, incluso il ricorso a un po' di blasone regionale» [21: 178].

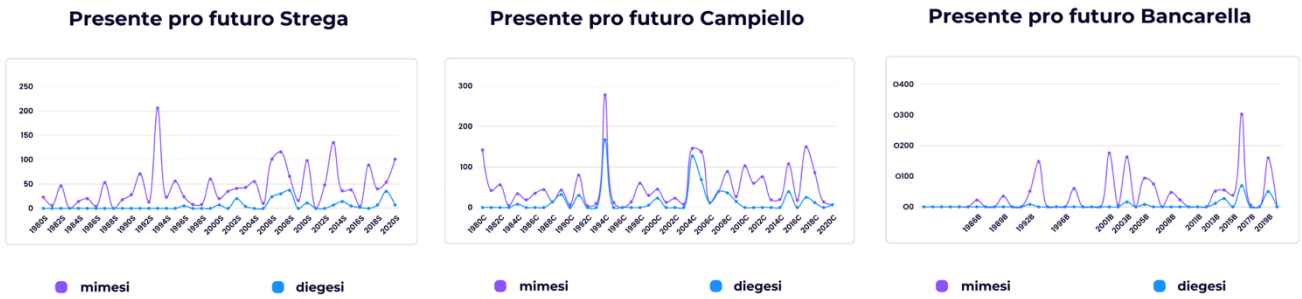


Figura 1. Differenziali delle occorrenze nei premi Strega, Campiello e Bancarella

Lo Strega registra nei suoi testi premiati 332 occorrenze di presente pro-futuro registrate, di cui 289 in mimesi e 43 in diegesi. In questo premio si può notare un lento aumento di forme del presente pro-futuro e quella che potrebbe essere una maggiore presenza di stili meno alti e sublimi a partire dal 2000 circa, mentre in precedenza sembrerebbe confermato il bando a un linguaggio più spontaneo. L'eccezione è il testo vincitore nel 1993, ovvero *Ninfa plebea* di Domenico Rea e questo sembrerebbe confermare un carattere dello Strega, peraltro condiviso con la maggior parte dei più famosi premi letterari stranieri: l'essere particolarmente sensibile alle pressioni commerciali esercitate dalle maggiori case editrici [21: 17]. Infatti, l'unicità dell'opera (pubblicata nel 1992) combacia con l'identità della casa editrice che l'ha portata sul mercato editoriale: è l'unica opera, delle 41 nel relativo premio, a non essere stata data alla stampa da un editore blasonato. L'opera di Rea è stata pubblicata dalla Leonardo editore, un marchio indipendente fino al 1993 e con una linea editoriale che non ha seguito i filoni tradizionali, ma ha preferito piuttosto la promozione della creatività all'interno delle pratiche di scrittura[1], fino a quando non ha incrociato le vicissitudini del Lodo Mondadori. Le altre opere che sembrano segnare delle controtendenze nel nuovo millennio sono quelle vincitrici delle edizioni 2006, 2007, 2010, 2013 e 2017; ovvero *Caos calmo* di Sandro Veronesi (2006S), *Come Dio comanda* di Nicolò Ammaniti (2007S), *Canale Mussolini* di Antonio Pennacchi (2010S), *Resistere non serve a niente* di Walter Siti (2013S) e *Le otto montagne* di Paolo Cognetti (2017S). Pennacchi e Siti hanno in comune il materiale su cui hanno costruito le loro narrazioni, cioè realtà sociolinguistiche a cui solitamente viene associata una minore vigilanza sulle forme. In Pennacchi è protagonista una famiglia di contadini della bassa padana, i Peruzzi, che muove verso il Lazio per la bonifica dell'omonimo canale. In Siti è protagonista una differente prospettiva generazionale all'interno della criminalità organizzata, dove un bagaglio di nuove competenze permette sì ai più giovani di allargare gli orizzonti criminali, ma non cancellerebbe l'oralità tipica dei contesti di socializzazione primaria di provenienza. Il carattere linguistico che i dati evincono in Veronesi, Ammaniti e Cognetti sembra coincidere con i risultati del *close reading* di Simonetti, per cui Veronesi e Ammaniti sarebbero accomunati da una scrittura "scenografica"

[1] «Leonardo editore srl (1986 -)», https://www.lombardiabeniculturali.it/archivi/soggetti-produttori/ente/MIDB001358/.

che pare *ab origine* confezionata per il salto transmediale e audiovisivo, come poi avvenuto [21: 138]. Anche Cognetti dimostrerebbe sintassi e ritmiche analoghe, ma oltre a questi aspetti, la sua scrittura risulterebbe essere anche la più distante dai modelli sublimi e più conservatori: «*Le otto montagne* […] risulta tra i cinque il romanzo più consapevole e forse, in fondo, il più orgoglioso della propria distanza culturale e linguistica da una certa tradizione italiana» [21: 140].
La doppia giuria del Campiello, per Simonetti, è nata per smarcarsi dalle logiche più meramente editoriali (Strega) e più meramente commerciali (Bancarella) e puntare sul valore della narrazione in sé, non sulle possibili auree di chi ha pubblicato e scritto. Dal 2000, la concorrenza avrebbe spinto il Campiello verso la prosa insieme commerciale e pseudo-impegnata che connoterebbe lo Strega [21: 105-106]. Delle 250 occorrenze registrate nel Campiello, 181 sono in mimesi e 69 in diegesi. Si possono notare alcune analogie lo Strega. Entrambi i grafici mostrano un picco nella prima metà degli anni Novanta (presagio degli imminenti cannibali?) e un lento aumento di forme del presente e di quella che potrebbe essere una maggiore presenza di stili meno alti e sublimi a partire dal 2000 circa. Nel Campiello il picco negli anni Novanta è *Sostiene Pereira* di Antonio Tabucchi (1994C). Qui è significativo notare come la granularità nei dati riesca a rilevare un una qualità diegetica riflessa nella complessità narratologica di un testo che sovrappone la voce di Pereira filtrata dal narratore a quella del narratore stesso. È anche significativa la granularità data dall'approccio nel differenziare i maggiori tre picchi degli anni 2000. Distingue in qualità diegetica *La barca nel bosco* di Paola Mastrocola (2004C) da *L'ultimo arrivato* di Marco Balzano (2015C) e *L'arminuta* di Donatella Di Pietrantonio (2017C), ma rileva anche una minima differenza diegetica tra 2015C e 2017C. I tre testi hanno in comune l'essere tutti autodiegesi condotte da pre-adolescenti e adolescenti. Da qui il primo, secondo e terzo posto per quel che riguarda gli anni 2000 (viene escluso 2005C di Roveredo perché l'averlo misurato non dividendolo in sotto-testi non permette un discorso organico e comparato). Ma i dati differenziano ulteriormente 2004C. Perché è una narrazione simultanea che riporta i fatti adolescenziali mentre avvengono per la maggior parte del libro. Infatti, 2004C presenta come narratore autodiegetico un adolescente (dai 13 ai 25 anni circa) e offre il lavoro autoriale nel filtrare una storia attraverso occhi e lingua di un ragazzino che da una piccola isola del Sud si trasferisce a Torino per studiare al liceo. 2015C ripresenta la figura del giovane emigrato da una più specificata Sicilia verso Milano, ma alla più tenera età di 9 anni. A differenza di 2004C, in 2015C il narratore autodiegetico a volte racconta in narrazione simultanea, altre invece compie un salto temporale narrando già da adulto e anziano, con *amarcord* e distanza che ne consegue. Infine, l'io narrante di 2017C è una ragazza che, nelle prime due parti del libro, cerca le risposte a quello che le è successo fra i 13 e 14, quando è stata rispedita indietro nella sua famiglia biologica, dopo che fin dai sei mesi era cresciuta in una famiglia adottiva di parenti. 2004C si differenzia per l'autodiegesi in narrazione simultanea che in 2015C e 2017C è più limitata. Inoltre, a dividere le tre opere nel frattempo c'è stata anche una certa Elena Ferrante che pur non vincendo nessuno di questi premi dal 2011riscuote successo con un io narrante che a distanza di tempo ripercorre le fasi adolescenziali della sua formazione (e in ambito di attribuzione autoriale se ne potrebbero ipotizzare due distinte usando questo approccio per darle la caccia).
Anche nel Bancarella le 198 occorrenze registrate, di cui 178 sono in mimesi e 20 in diegesi, suggeriscono analogie. Il più alto picco, *La ragazza di fronte* di Margherita Oggero (2016B), è un romanzo che segue due personaggi principali, lei archivista torinese e lui un meridionale che dal Sud ha fatto base a Torino, impiegando due eterodiegesi che si avvicendano in narrazioni simultanee. I dati sembrerebbero collegare queste caratteristiche di 2016B a quelle analoghe già viste in 2004C e 2015C. I dati, infine, segnano altri tre picchi: *La gita a Tindari* di Andrea Camilleri (2001B), *Amiche di salvataggio* di Alessandra Appiano (2003B) e *Il ladro gentiluomo* di Alessia Gazzola (2019B). Anche qui l'approccio permette di inferire e notare analogie: 2001B è un giallo; 2003B è un *chick lit*; 2019B un *chick lit* ibridato di giallo.

5. CONCLUSIONI E IPOTESI DI SVILUPPI FUTURI

Se notiamo i diversi rapporti tra il numero totale delle occorrenze verbali di ogni premio, le percentuali delle occorrenze del presente pro-futuro nei tre premi e il differenziale dei dati, il Bancarella segna allo stesso tempo le più alte percentuali di uso del presente pro-futuro (0,062%) e di impieghi mimetici (89,90% in mimesi e 10,10% in diegesi). Il Campiello e lo Strega hanno valori simili sull'uso generale del presente pro-futuro (Campiello 0,041% e Strega 0,039%), ma differenziali che li distinguono nettamente sull'asse diegetico-mimetico. Il Campiello è al primo posto per gli impieghi diegetici (27,6% in diegesi e 72,4% in mimesi) e doppia l'impiego diegetico del presente pro-futuro nello Strega (12,95% in diegesi e 87,05% in mimesi).
Se continuiamo a ipotizzare che una maggiore percentuale generale di impieghi del presente coincida con una lingua più spontanea e meno vigilata mentre una maggiore percentuale di impieghi diegetici coincida con una complessità delle strutture narrative, troviamo il premio popolare Bancarella aver apprezzato di più la spontaneità linguistica, il Campiello aver apprezzato di più la complessità narrativa e lo Strega equidistante: essersi tenuto a debita distanza da complessità narrativa e spontaneità linguistica, come sostenuto nello studio di Simonetti. È un risultato significativo che già con pochi

dati l'approccio sia in grado di raccogliere al contempo indicazioni diegetiche sulle complessità narrative e mimetiche su tipo di lingua e milieu finzionali. Infine, l'approccio proposto appare promettente perché già con un basso output di dati riesce a far inferire ed emergere analogie tra realtà linguistico-narrative effettivamente presenti nei testi letterari, ma diversamente perse limitando lo studio nell'attuale invalso del *Distant reading*.

L'approccio, ovviamente, necessita di ulteriori studi e conferme con altri *corpora*, ma è già possibile avanzare linee di sviluppi futuri a cui dà base e strumenti epistemici. Uno degli sviluppi futuri della ricerca riguarda la valutazione dei dati sotto una cornice teorica arricchita dalla *Game Theory*. In questo caso si ipotizza che la scrittura diegetica e la scrittura mimetica siano due *player* per come intesi nella *Game Theory* e che il genere letterario sia la risultante dei differenti rapporti che tra essi si possono instaurare. Tali rapporti sono misurabili con analisi dei sentimenti, modellazione degli argomenti, liste di vocaboli e analisi differenziali. Infatti, una applicazione teorica della *Game Theory* al rapporto tra le scritture diegetiche e mimetiche risulta estremamente interessante per tre motivi.

Il primo motivo è che negli studi letterari la *Game Theory* è stata applicata solo in ambito di scrittura mimetica, nel senso che è stata fatta la cosa più ovvia: analizzare il comportamento dei personaggi alla luce delle categorie della teoria dei giochi [4: 1-26]. Ma non è stato ipotizzato un profilo per le scritture diegetica e mimetica in quanto due *player* né un rapporto tra esse in quanto *player*.

Il secondo motivo è che le varie categorie teoriche previste dalla teoria dei giochi sembrano coincidere con i rapporti che sussistono tra le scritture diegetiche e mimetiche, realizzati, di solito, per differenziare generi letterari e specifici espedienti letterari. Così, la categoria del gioco cooperativo sembra riflettere il rapporto tra le scritture diegetiche e mimetiche nei generi letterari del romanzo storico, della fiaba, dell'epica, dove la comunanza di interessi (la cooperazione) risiede in narrazioni nelle quali, tendenzialmente, non si avrà una delle due scritture che contraddice l'altra, piuttosto appunto appaiono coordinate a un comune fine narrativo. Dall'altra parte, la categoria del gioco non cooperativo sembra riflettere il rapporto tra le scritture diegetiche e mimetiche nei generi letterari del thriller e dell'horror, dove il non coordinamento, lo scarto tra le due scritture è atto alla creazione della necessaria suspense. La categoria dei giochi ripetuti nel tempo e i loro continui *payoff* sembrano riflettere il rapporto tra le scritture diegetiche e mimetiche che connotano le serialità letterarie e i godimenti cognitivi dati dall'isomorfismo. La categoria dei giochi a informazione perfetta sembrano riflettere i tipici rapporti tra le scritture diegetiche e mimetiche delle narrazioni con narratore onnisciente; dall'altra parte, la categoria dei giochi a informazione imperfetta sembrano riflettere i tipici rapporti delle scritture diegetiche e mimetiche nelle narrazioni con narratore inattendibile. La categoria dei giochi finiti sembrano riflettere i tipici rapporti tra le scritture diegetiche e mimetiche nei generi letterari giallo e rosa. Infine, la categoria dei giochi a somma zero coincide con i tipici rapporti della scrittura mimetica impiegata nella tragedia, il genere letterario in cui maggiormente è stata applicata la *Game Theory*. È questa una ipotesi che chiede approfondimenti, ma che già allo stadio iniziale sembra promettente, anche quando si considera il fatto che la rivoluzione storico-letteraria del metateatro introdotta da Pirandello, in questa prospettiva, risulta appunto dall'aggiunta del nuovo *player* autoriale, laddove per millenni c'era stata la sola dimensione mimetica. E anche nella narrativa le infrazioni tra realtà e finzione tipiche dei testi metaletterari scaturiscono dall'aggiunta del nuovo *player* autoriale, laddove per millenni erano prevalse soprattutto le sole contrapposizioni tra le scritture diegetiche e mimetiche. In questa ipotesi i diversi rapporti (e infrazioni) tra le scritture diegetiche e mimetiche che distinguono i generi letterari assumono l'immagine di concerti ad archi; un'immagine rafforzata dall'elemento *game* della teoria che quando messo in azione esprime in diverse lingue il significato di "suonare" oltre che di "giocare": inglese *to play*, francese *jouer*, tedesco *spielen*, russo играть (*igrat'*).

Il terzo motivo è che la *Game Theory* è nata in ambito matematico, ma ha avuto le sue migliori applicazioni nelle scienze sociali prima di tutto e poi nell'evoluzionismo e nel cognitivismo: tre ambiti teorico-applicativi verso cui il collegamento del *Distant reading* viene sempre più indirizzato e considerato necessario.

BIBLIOGRAFIA

[1] Aristotele. *Poetica*. (a cura di) M. Valgimigli. Bari: Laterza, 1964.

[2] Berruto, Gaetano. *Sociolinguistica dell'italiano contemporaneo*. Roma: La Nuova Italia Scientifica, 1987.

[3] Bertinetto, Pier Marco. «Metafore tempo-aspettuali». *Linguistica* 32, fasc. 2 (1992): 89-106.

[4] Brams, J. Steven. *Game Theory and The Humanities. Bridging two Worlds*. 2011a ed. Cambridge-London: The MIT Press, 2011.

[5] Ciotti, Fabio. «Distant reading in literary studies: a methodology in quest of theory». *Testo e Senso* 23 (2021): 195-213. https://testoesenso.it/index.php/testoesenso/article/view/509.

[6] De Beaugrande, Robert-Alain, e Dressler Wolfgang. *Introduction to Text Linguistics*. New York: Routledge, 1981.

[7] Eco, Umberto. *Kant e l'ornitorinco*. Milano: Bompiani, 1997.

[8] Firth, John Rupert. «The technique of semantics». *Transactions of the Philological Society* 34, fasc. 1 (1935): 36-72.

[9] Gavin, Micheal. *Literary Mathematics. Quantitative Theory for Textual Studies*. Stanford: Stanford University Press, 2022.

[10] Genette, Gérard. *Figures* II. Paris: Seuil, 1969.

[11] Genette, Gérard. *Figures* III. Paris: Seuil, 1972.

[12] Harris, S. Zellig. «Distributional Structure». *Word* 10, fasc. 2-3 (1954): 146-162.

[13] Lorenzetti, Luca. *L'italiano contemporaneo*. Roma: Carocci, 2002.

[14] Moretti, Franco. «'Operationalizing'». *New Left Review* 84 (novembre 2013): 103-119.

[15] Mueller, Martin. «Shakespeare His Contemporaries: collaborative curation and exploration of Early Modern drama in a digital environment». *Digital Humanities Quarterly* 8, fasc. 3 (2014). http://www.digitalhumanities.org/dhq/vol/8/3/000183/000183.html.

[16] Patota, Giuseppe. *Grammatica di riferimento dell'italiano contemporaneo*. Milano: Garzanti, 2010.

[17] Piper, Andrew. «Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel». *New Literary History* 46, fasc. 1 (2015): 63-98.

[18] Platone. *La Repubblica*. (a cura di) G. Lozza. Milano: Mondadori, 1990.

[19] Sabatini, Francesco. «"Rigidità-esplicitezza" vs "elasticità-implicitezza": possibili parametri massimi per una tipologia dei testi». (a cura di) G. Skytte e F. Sabatini, 141-172. Copenhagen: Museum Tusculanum Press, 1999.

[20] Serianni, Luca. *Grammatica italiana. Italiano comune e lingua letteraria*. Torino: UTET Università, 2006.

[21] Simonetti, Gianluigi. *Caccia allo Strega. Anatomia di un premio letterario*. Milano: Nottetempo, 2023.

[22] Werlich, Egon. *A text grammar of English*. Heidelberg: Quelle und Meyer, 1976.

L'impiego dell'intelligenza artificiale per la ricostituzione delle aggregazioni archivistiche e l'arricchimento dei metadati negli archivi digitali

Stefano Allegrezza
Università degli Studi di Bologna, Italia - stefano.allegrezza@unibo.it

ABSTRACT

Il contributo intende presentare i primi risultati di uno studio condotto nell'ambito del progetto internazionale InterPARES Trust AI ed intitolato «The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas». L'obiettivo generale di questo studio è quello di indagare la capacità dell'intelligenza artificiale di supportare la creazione (o la ri-costituzione) di aggregazioni archivistiche per risolvere il problema della presenza di documenti non aggregati, non ordinati o de-contestualizzati (sia nella fase corrente che in quella semi-corrente dell'archivio) che si presenta in molte situazioni. Spesso, infatti, sia nelle amministrazioni pubbliche che nelle aziende, i documenti non vengono né classificati né fascicolati; oppure, le aggregazioni documentali vengono formate ma in modo non corretto. Inoltre, non di rado i metadati – che sono necessari per garantire l'autenticità, l'affidabilità e la ricercabilità – non vengono correttamente individuati ed associati ai documenti. In questo modo l'archivio dell'organizzazione non viene formato correttamente, e ciò costituisce una grave criticità perché fa sì che ci sia un numero incontrollato di documenti non ordinati, mal collocati e quindi difficili da trovare. Purtroppo, nonostante i progressi compiuti dalle tecnologie informatiche per fornire un aiuto nelle attività di gestione documentale, bisogna riconoscere che gli attuali prodotti software sono in grado di fornire un supporto molto limitato a questo tipo di esigenze. Tuttavia, le tecniche di intelligenza artificiale sembrare promettere grossi passi avanti in questo campo. Lo studio che si intende presentare si pone proprio l'obiettivo di fornire una risposta alle seguenti domande di ricerca: gli strumenti di intelligenza artificiale possono aiutare a creare le aggregazioni documentali quando queste non sono mai state formate o a ri-crearle quando erano state formate ma sono andate perdute? Possono fornire un valido aiuto nell'individuazione di metadati e nella associazione ai documenti relativi?

PAROLE CHIAVE

Intelligenza artificiale; archivi; fascicoli; metadati; aggregazioni documentali.

1. INTRODUZIONE

La formazione dei fascicoli e delle aggregazioni documentali rientra tra le operazioni archivistiche strategiche ai fini della corretta formazione dell'archivio. Infatti «la decisione di aggregare un nuovo documento archivistico ad un fascicolo già aperto oppure di aprire un nuovo fascicolo in seguito all'acquisizione di un certo documento archivistico consente l'ordinato stratificarsi della produzione documentaria di un ente nel corso della sua concreta e quotidiana attività amministrativa all'interno di un quadro logico astratto, cioè il titolario di classificazione» [4]. Tale operazione è fondamentale anche nel caso degli archivi nativi digitali, tanto che, con riferimento al contesto italiano, l'articolo 64, comma 4, del "Testo Unico sulla documentazione amministrativa" prevede che «le amministrazioni determinano autonomamente e in modo coordinato per le aree organizzative omogenee, le modalità di attribuzione dei documenti ai fascicoli che li contengono e ai relativi procedimenti, definendo adeguati piani di classificazione d'archivio per tutti i documenti, compresi quelli non soggetti a registrazione di protocollo»[1]. Tale previsione viene ripresa anche dall'art. 71 del "Codice dell'amministrazione digitale" il quale stabilisce che «la pubblica amministrazione titolare del procedimento raccoglie in un fascicolo informatico gli atti, i documenti e i dati del procedimento medesimo da chiunque formati»[2]. Da ultimo, anche le "Linee Guida sulla formazione, gestione e conservazione dei documenti informatici" pubblicate dall'Agenzia per l'Italia Digitale (AgID) stabiliscono che le Pubbliche Amministrazioni gestiscono «i flussi documentali mediante fascicoli informatici predisposti secondo il piano di classificazione e relativo piano di organizzazione delle aggregazioni documentali» [1].

[1] Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445. "Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa". (GU n.42 del 20-02-2001 - Suppl. Ordinario n. 30).
[2] Decreto Legislativo 7 marzo 2005, n. 82. "Codice dell'amministrazione digitale". (GU n.112 del 16-05-2005 - Suppl. Ordinario n. 93). Art. 41 «Procedimento e fascicolo informatico».

La creazione dei fascicoli, quindi, non è soltanto una buona prassi dal punto di vista archivistico, ma è anche un obbligo normativo. Purtroppo, la situazione che si riscontra diffusamente, sia nelle realtà pubbliche che in quelle private, è una insufficiente attenzione nei confronti di questa operazione, tanto necessaria – in quanto la sua mancanza porta alla formazione di archivi disordinati, caotici e molto difficili da riordinare in una fase successiva – quanto trascurata. In molte amministrazioni pubbliche i documenti non sono classificati né aggregati in fascicoli. In altri casi, le aggregazioni documentali non sono create correttamente, con il risultato della presenza di un numero incontrollato di documenti non ordinati, non collocati nel fascicolo corretto e difficili da trovare. In molti casi mancano i metadati, «utilizzati per finalità molteplici: ricerca, gestione e localizzazione, selezione, interoperabilità» [5] e necessari per garantire l'affidabilità, l'autenticità, la qualità.

Nonostante i progressi compiuti dalle varie tecnologie a supporto della gestione dei documenti, la disponibilità di software per questo tipo di attività rimane ad oggi ancora molto limitata. Il motivo per cui l'operazione di formazione delle aggregazioni documentali è così trascurata è riconducibile a diverse motivazioni, come la mancanza di conoscenze archivistiche, la poca consapevolezza della sua importanza, la mancanza di volontà, o, infine, la mancanza di tempo. Tutto ciò ha delle conseguenze rilevanti non solo in termini di perdita di efficienza dell'azione amministrativa, causata dalla difficoltà di gestire il flusso procedimentale dal momento che i documenti ad esso relativi non si trovano riuniti all'interno della stessa aggregazione documentale, ma anche dal punto di vista della preservazione della memoria, dal momento che un archivio del genere sarà anche difficilmente fruibile da parte delle generazioni future di ricercatori e studiosi. Fino ad oggi la soluzione a questo problema è stata cercata soprattutto nell'organizzazione di interventi mirati di formazione con lo scopo di sensibilizzare gli operatori degli uffici sull'importanza dell'operazione di fascicolazione, ma questi interventi non sempre si sono rivelati efficaci e spesso non hanno sortito l'effetto desiderato.

Viene naturale, allora, chiedersi se le tecnologie di intelligenza artificiale (IA) possano essere utili per una gestione efficiente, automatica o semi-automatica, delle aggregazioni documentali o per l'arricchimento dei metadati, svolgendo così quei compiti che l'essere umano non riesce a compiere.

Un caso paradigmatico è quello relativo alla gestione dell'archivio di posta elettronica, che è una delle attività più dispendiose in termini di tempo sia nel settore pubblico che in quello privato. La dottrina archivistica insegna che le email ricevute ed inviate, così come in generale tutta la documentazione, andrebbero ordinate all'interno di una struttura articolata in fascicoli e sotto-fascicoli ('cartelle' e 'sotto-cartelle' nel gergo informatico) [3], create in relazione ai vari 'affari' trattati, ai vari progetti, alle varie attività, oppure in relazione ai vari corrispondenti; l'articolazione di questa struttura «non dovrebbe essere né eccessivamente dettagliata (perché altrimenti l'assegnazione di ciascuna email nella cartella di competenza richiederebbe troppo tempo) né eccessivamente generale (perché altrimenti perderebbe di efficacia); di solito un numero di livelli compreso tra due e tre è la scelta migliore» [2]. Quello che andrebbe assolutamente evitato è di lasciare tutte le e-mail nelle cartelle della "Posta in arrivo" e della "Posta inviata" dal momento che così facendo ben difficilmente sarà possibile mettere ordine in quella massa disordinata e confusa di e-mail, che spesso raggiunge dimensioni importanti (nell'arco di una vita può facilmente raggiungere la dimensione di decine di migliaia di messaggi e a volte anche di più). Purtroppo, riuscire a formare un archivio organizzato in questo modo non è semplice dal momento che aggregare nelle relative cartelle – del client di posta o dell'applicazione webmail – la quantità spesso eccessiva di e-mail che quotidianamente vengono ricevute ed inviate richiederebbe di dedicare un tempo non indifferente a questa attività. Questo è probabilmente il motivo per cui la maggior parte delle persone – salvo casi particolari – lascia tutte le e-mail nelle cartelle della "posta in arrivo" e della "posta inviata" creando così un archivio sostanzialmente privo di aggregazioni documentali dove rintracciare una mail è ben difficile (se non affidandosi alle funzioni di ricerca, che però richiedono che la mail ricercata abbia almeno un oggetto ben definito)[4] e dove nel futuro sarà molto difficile condurre studi e ricerche.

[3] Le 'cartelle' sono gli equivalenti informatici dei fascicoli archivistici. Secondo l'Allegato 5 "Metadati" alle *Linee guida sulla formazione, gestione e conservazione di documenti informatici* dell'Agenzia per l'Italia Digitale [1] esistono quattro tipologie di fascicoli: fascicolo per *affare*: conserva i documenti relativi a una competenza non proceduralizzata, ma che nella consuetudine amministrativa la PA deve concretamente portare a buon fine. Il fascicolo per affare ha una data di apertura e una durata circoscritta; fascicolo di *persona fisica* o *giuridica*: comprende tutti i documenti, anche con classifiche diverse, che si riferiscono a una persona (fisica o giuridica). Quasi sempre i fascicoli intestati alle persone restano correnti per molti anni, costituendo serie aperte; fascicolo *per attività*: comprende i documenti prodotti nello svolgimento di un'attività amministrativa semplice che implica risposte obbligate o meri adempimenti, per la quale quindi non è prevista l'adozione di un provvedimento finale. Ha in genere durata annuale; fascicolo per *procedimento amministrativo*: conserva una pluralità di documenti che rappresentano azioni amministrative omogenee e destinate a concludersi con un provvedimento amministrativo. I fascicoli sono cinque se si considerano distinti i fascicoli di persona fisica da quelli di persona giuridica.

[4] Si potrebbe obiettare che le funzioni di ricerca consentono di effettuare ricerche di tipo *full-text* – che prendono in considerazione anche il corpo del messaggio – e, combinate con i filtri di ricerca (mittente, destinatario, data, presenza di allegati, ecc.). potrebbero consentire di trovare una determinata e-mail. Su questo punto occorre fare un paio di considerazioni. Innanzitutto, le ricerche di questo tipo non sempre consentono di individuare l'e-mail che interessa perché spesso restituiscono un numero molto elevato di risultati, specialmente

Anche in questo caso viene spontaneo chiedersi se via siano applicazioni basate sull'intelligenza artificiale in grado di riunire all'interno di una aggregazione documentale le e-mail scambiate con un certo corrispondente o relative ad un determinato 'affare', delegando, cioè, all'intelligenza artificiale lo svolgimento di quelle operazioni che l'uomo non riesce a svolgere.

2. METODOLOGIA

Per ovviare alle criticità appena evidenziate, all'interno del progetto di ricerca internazionale "InterPares Trust AI"[5] il gruppo di lavoro denominato CU05 ("Creation and Use 05") è stato incaricato di condurre uno studio per indagare se le tecniche di intelligenza artificiale possano essere d'aiuto nelle operazioni di formazione delle aggregazioni documentali e di individuazione dei metadati. In particolare, questo studio mira a valutare se le tecnologie di IA esistenti possono ristabilire il legame archivistico tra una moltitudine di documenti decontestualizzati e integrare schemi di metadati incompleti.

L'indagine ha avuto inizio con un censimento delle aziende che dichiarano di sviluppare prodotti basati sull'intelligenza artificiale applicata ai settori della gestione documentale e che sono rilevanti nell'ambito dello studio. Questa prima fase, che si è svolta tra febbraio e giugno 2022, ha portato all'identificazione di un elenco iniziale – non esaustivo – di circa 300 aziende. L'elenco è stato costruito sia mediante ricerche dirette svolte su Internet utilizzando parole chiave[6] e stringhe di testo, sia grazie alle risorse e conoscenze messe a disposizione da alcuni professionisti [cfr. sez. "Ringraziamenti" alla fine del contributo]. Le caratteristiche dei software di IA sono state analizzate in prima battuta in base alle informazioni disponibili sui rispettivi siti web (ad esempio, se l'azienda dichiara che la propria applicazione di intelligenza artificiale è applicabile al settore della gestione dei documenti o a quello della conservazione degli archivi, anche se in alcuni casi non vi è una dichiarazione esplicita, ma la si può solo intuire dal contenuto del sito web).

Successivamente, dall'elenco iniziale è stato selezionato un elenco ristretto di 100 aziende – la cui distribuzione geografica è riportata in Figura 1 – identificate come potenzialmente molto interessanti per gli scopi dello studio e che sarebbero state ulteriormente vagliate. Infine, l'elenco delle 100 aziende è stato ulteriormente ridotto andando a selezionare quelle aziende che – sulla base del portafoglio clienti, del coinvolgimento diretto nel settore della gestione documentale, della conformità ai quadri normativi e agli standard relativi alla gestione degli archivi e dei documenti e agli standard di settore e della reputazione generale dell'azienda – meglio rispondevano agli scopi dello studio. Da questa ulteriore fase di 'scrematura', che si è svolta tra il giugno e l'agosto 2022, è emerso un elenco di 28 aziende – potenzialmente le più interessanti, ma evidentemente l'elenco potrebbe non essere esente da errori e fraintendimenti – alle quali è stata inviata una lettera ufficiale (in inglese, spagnolo o portoghese, a seconda della lingua preferita dall'azienda) con l'invito a partecipare all'indagine.

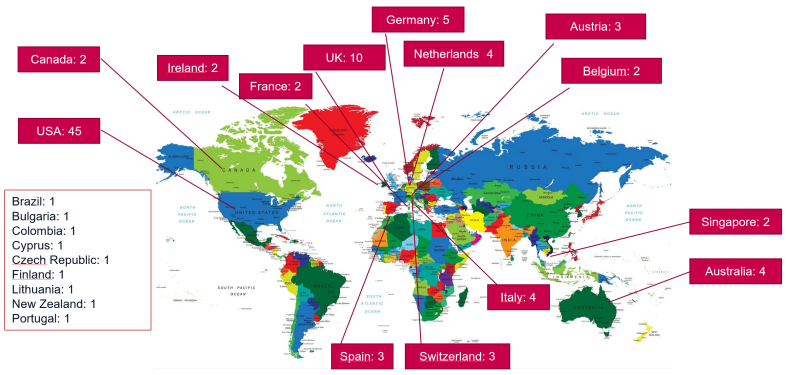


Figura 1. Distribuzione geografica delle 100 aziende selezionate

quando i parametri di ricerca non sono ben definiti (ad esempio, non ci si ricorda che è il mittente, quale fosse l'oggetto dell'e-mail, o il periodo in cui è stata ricevuta o inviata, ecc.). In secondo luogo, quando si parla di ricerca archivistica non è sufficiente il ritrovamento di una singola e-mail ma occorre ritrovare l'intero 'carteggio' elettronico intercorso con un certo corrispondente o relativamente ad un certo 'affare': questo tipo di risultato lo si ottiene solo con una corretta organizzazione dell'archivio di posta elettronica.

[5] InterPARES Trust AI. https://interparestrustai.org.

[6] Ad esempio, sono state effettuate ricerche utilizzando stringhe di ricerca come "artificial intelligence", "records", "documents", "information extraction", "archival bond"; oppure "artificial intelligence", "records", "document classification", "file plan", "archives"; o, ancora, "artificial intelligence", "records management", "document classification", "recordkeeping", "archives". La ricerca è stata condotta anche utilizzando altre lingue oltre a quella inglese.

Allo scopo di raccogliere in maniera sistematica le informazioni sugli applicativi di intelligenza artificiale fornite dalle aziende intervistate, il gruppo di lavoro ha preparato un dettagliato questionario in lingua inglese, necessario per un'adeguata valutazione delle applicazioni destinate a supportare la ricostituzione delle aggregazioni archivistiche e l'arricchimento dei metadati. Il questionario è composto da quattro sezioni, ciascuna con diverse domande aperte.
La sezione I (*Risultati conseguiti*) si concentra sui risultati conseguiti dalle aziende, in particolare sulle caratteristiche delle applicazioni, sulle piattaforme di sviluppo, sulle caratteristiche principali e sui punti di forza, sugli aspetti da migliorare, sugli sviluppi futuri, nonché, infine, sulla conformità agli standard archivistici e di gestione dei documenti. La sezione II (*Capacità specifiche per i sistemi di gestione documentale e di posta elettronica*) si occupa delle funzionalità specifiche delle applicazioni per la gestione dei documenti, compresa quelli dei sistemi di posta elettronica. Le domande riguardano l'automazione di diverse attività di gestione dei documenti, come l'archiviazione dei documenti in fascicoli secondo uno schema di classificazione dei documenti, la valutazione e lo scarto dei documenti secondo un piano di conservazione, l'estrazione di metadati e l'indicizzazione dei documenti per fornire informazioni su aggregazioni correlate. La sezione III (*Tecnologie e metodi utilizzati nelle applicazioni di IA*) analizza le tecnologie utilizzate nelle applicazioni di IA, come i modelli, le strategie e le tecniche di apprendimento automatico, nonché gli elementi dei documenti o dei metadati che l'applicazione prende in considerazione per prendere decisioni e fare inferenze. Infine, la sezione IV (*Controlli di audit – indicatori di prestazione*) si concentra sulle verifiche e sugli indicatori di prestazione per misurare la percentuale di successo e di insuccesso delle applicazioni e le eventuali distorsioni.
Alla lettera di invito hanno risposto positivamente 13 aziende (vd. Fig. 2) che hanno accettato di partecipare all'indagine e con ciascuna delle quali è stato organizzato un incontro on-line per illustrare il questionario e le sue finalità. All'incontro, oltre al gruppo di lavoro, hanno partecipato i referenti delle aziende intervistate: di solito hanno partecipato coloro che si occupavano della gestione documentale, gli ingegneri informatici e gli archivisti (se presenti); si è privilegiata la partecipazione del personale tecnico e si è cercando di evitare la partecipazione dei responsabili commerciali (che avrebbero probabilmente esaltato le caratteristiche delle loro soluzioni fornendo così informazioni distorte). Successivamente, le aziende hanno compilato il questionario che è stato reso disponibile su Google Forms fino al mese di febbraio 2023.

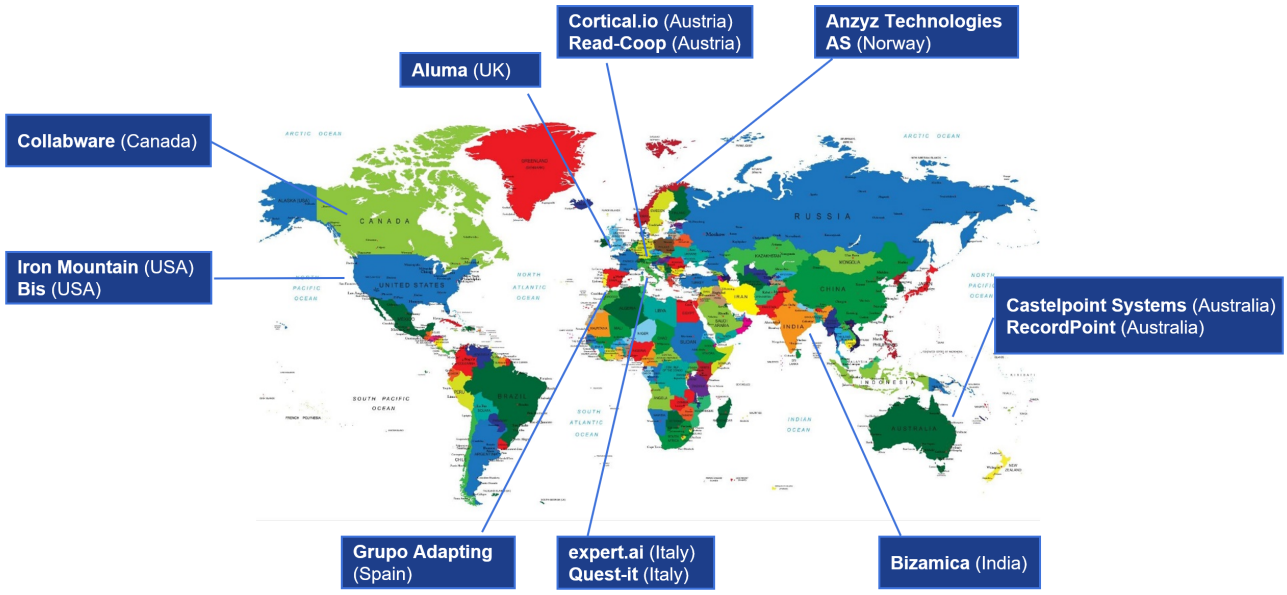


Figura 2. Le 13 aziende che hanno risposto al questionario

3. RISULTATI OTTENUTI

Il gruppo di lavoro ha preso in esame le risposte ottenute e ha elaborato un report che è stato pubblicato nel mese di novembre 2023 sul sito del progetto internazionale [3]. Non essendo possibile, per mancanza di spazio, discutere in questa sede le risposte ottenute nel questionario, si riportano alcune conclusioni di carattere generale.
Innanzitutto, tutte le aziende intervistate hanno sviluppato soluzioni basate su tecnologie di IA per l'indicizzazione, la categorizzazione o la classificazione di documenti strutturati, semi-strutturati e non strutturati, sulla base di tecniche di apprendimento automatico e di estrazione automatica dei dati.

Per quanto riguarda la possibilità di creazione o ricostituzione di aggregazioni documentali basata su tecniche di intelligenza artificiale, le risposte ottenute non sono molto incoraggianti, in quanto è emerso come questa possibilità sia limitata a casi molto specifici: ad esempio, solo per determinate tipologie di documenti, oppure quando vi sono almeno delle informazioni di base sulla struttura delle aggregazioni, o, ancora, quando l'aggregazione automatica viene solamente proposta e necessita di una fase di convalida da parte dell'utente.

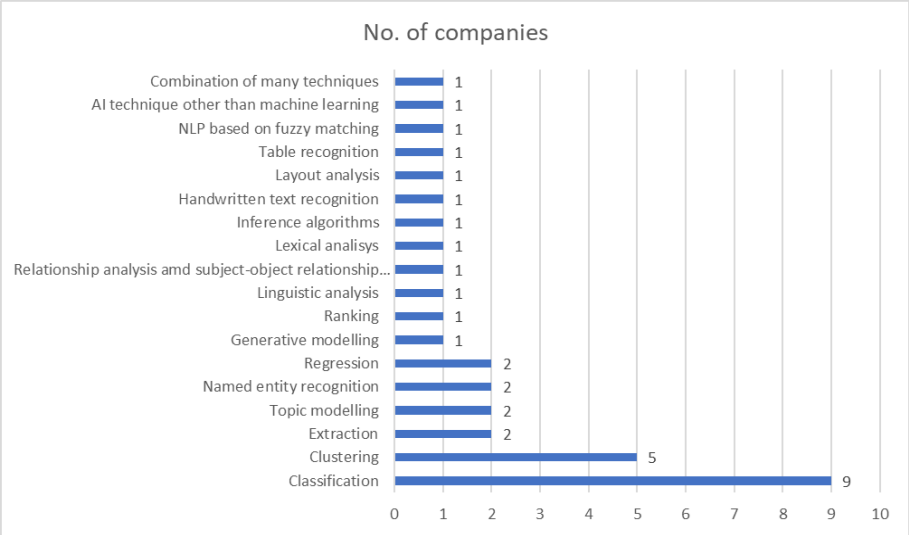


Figura 3. Le tecniche di intelligenza artificiale dichiarate dalle aziende che hanno risposto al questionario

L'indagine conteneva alcune domande sulle soluzioni tecniche adottate dalle aziende, e più precisamente sui modelli di analisi e sui tipi di tecniche utilizzate nei prodotti, sulle strategie di formazione, sugli elementi informativi elaborati dalle piattaforme e sulle caratteristiche degli ecosistemi informatici di cui i prodotti basati sull'IA hanno bisogno per funzionare correttamente. Per quanto riguarda i modelli di analisi, il panorama risulta molto variegato: le aziende dichiarano di utilizzare complessivamente 24 modelli diversi. *Neural networks* (7 volte) e *support vector machines* (4 volte) sono i modelli più citati (va notato che in alcuni casi le risposte sono state piuttosto generiche). Per quanto riguarda le tecniche utilizzate, sono state citate complessivamente ben 18 diverse tipologie (vd. Fig. 3); la *classification* (9 volte) e il *clustering* (5 volte) sono state le risposte più comuni (questo era prevedibile, dato che le aziende sono state selezionate per la loro esperienza nella gestione dei documenti).

Infine, per quanto riguarda le strategie di apprendimento per i prodotti basati sull'IA, l'indagine ha mostrato una situazione mista (vd. Fig. 4): 11 aziende utilizzano l'apprendimento *supervised*, 6 aziende l'apprendimento *unsupervised*, 4 aziende l'apprendimento *semi-supervised*, 2 aziende l'apprendimento *auto-supervised* e 2 aziende l'apprendimento *rule-based*. Naturalmente, diverse aziende utilizzano più di un tipo di strategie di apprendimento.

Per quanto riguarda le caratteristiche degli ecosistemi informatici necessari per far funzionare i prodotti basati sull'IA, la maggior parte dei prodotti che sono stati esaminati possono funzionare in ogni sistema operativo (Windows, Linux, Mac, ecc.) e interagire con un numero molto elevato di piattaforme e applicazioni grazie ad API e connettori personalizzati, il cui sviluppo richiede, tuttavia, investimenti in termini di risorse finanziarie ed umane. Infine, quasi tutti i prodotti basati sull'intelligenza artificiale recensiti possono essere distribuiti e funzionare sia in locale che su cloud.

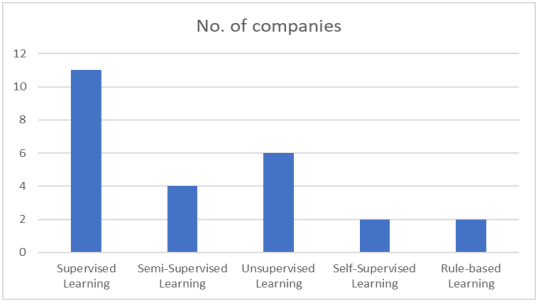


Figura 4. Le strategie di apprendimento dichiarate de dalle aziende che hanno risposto al questionario

4. CONCLUSIONI E ULTERIORI SVILUPPI

In conclusione, quello che è emerso dallo studio è che allo stato attuale, nonostante i promettenti annunci pubblicitari presenti su molti dei siti web di aziende che sviluppano di applicazioni *AI-based* per la gestione dei documenti, la ricostituzione del vincolo archivistico – elemento fondamentale delle aggregazioni documentali [6] – quando non è stato mai presente o è andato perduto non è un'operazione semplice né realizzabile solamente affidandosi a tali applicazioni; è sempre necessaria la disponibilità di ulteriori informazioni sulla struttura delle aggregazioni e/o un aiuto significativo da parte di operatori umani. Tuttavia, in molti casi le aziende hanno dichiarato che queste funzionalità non sono ancora disponibili perché ancora in fase di sviluppo, ma potrebbero esserlo a breve. A questo proposito, non è possibile sapere se nel futuro si potranno ottenere risultati migliori, ma certamente ciò sarà ottenuto solo al prezzo di maggiori investimenti, e che, al momento, non sembrano essere sostenuti da una corrispondente richiesta da parte del mercato.

Un altro elemento importante che è stato messo in luce dallo studio è la 'sfida' della terminologia. Quando si interagisce con le aziende che sviluppano prodotti basati sull'intelligenza artificiale, si ha normalmente a che fare con professionisti informatici, che il più delle volte attribuiscono a determinati termini un significato diverso da quello attribuito loro dagli archivisti. Ad esempio, questo è il caso di parole come 'classificazione', 'indicizzazione', 'clustering' che hanno un significato diverso nei due domini. Allo stesso modo, espressioni come 'Intelligent Document Processing', che si trovano sui siti di alcune aziende, non hanno nulla a che fare con il documento, il suo trattamento e, in ultima analisi, con l'intelligenza archivistica. Quindi, affinché il confronto con le aziende risulti davvero proficuo, è bene dedicare del tempo all'inizio a chiarire i concetti di base e prestare molta attenzione affinché essi vengano interpretati correttamente.

In sintesi, lo studio testimonia che, almeno al momento, la complessità delle funzioni archivistiche non può essere delegata completamente ad un approccio automatico basato sulle tecniche di intelligenza artificiale; tutt'al più può essere supportata da queste tecnologie, ma sempre sotto la supervisione o l'intermediazione di archivisti ed utenti.

Non è possibile affermare, senza ulteriori analisi e casi di studio, quale grado di intermediazione professionale sarà necessario in futuro, almeno in una prospettiva a breve o medio termine. Sono necessarie ulteriori indagini per valutare e misurare la qualità, l'affidabilità e l'accuratezza dei nuovi strumenti di IA e per verificare quanto sia vera la promessa che viene da più parti annunciata, ovvero quella di poter svolgere automaticamente i compiti di aggregazione dei documenti e di identificazione e inserimento dei metadati, anche sostituendo gli operatori umani. Per questi motivi il gruppo di lavoro CU05 sta proseguendo l'attività di ricerca attraverso l'analisi di alcuni casi di studio proprio con l'obiettivo di testare in situazioni concrete le effettive possibilità offerte dalle soluzioni di intelligenza artificiale che evolvono costantemente e incrementano giorno dopo giorno le proprie capacità.

5. RINGRAZIAMENTI

Ringrazio i colleghi Maria Mata Caravaca, Massimiliano Grandi, Mariella Guercio e Bruna La Sorda che, insieme a me, hanno contribuito alla realizzazione dello studio.

Ringrazio anche Alan Pelz-Sharpe, Andrew Warland, James Lappin, Jenny Bunn e Paul Young per l'indispensabile aiuto fornito nell'individuazione del primo elenco di 300 aziende informatiche che sviluppano prodotti di intelligenza artificiale applicata al settore della gestione documentale.

BIBLIOGRAFIA

[1] Agenzia per l'Italia Digitale. *Linee Guida sulla formazione, gestione e conservazione dei documenti informatici*, 2021.

[2] Allegrezza, Stefano. *La conservazione degli archivi di posta elettronica*. Torre del Lago Puccini: Civita Editoriale, 2022.

[3] Allegrezza, Stefano, Guercio, Mariella, Caravaca, Maria Mata, Grandi, Massimiliano, e La Sorda, Bruna. «The role of AI in identifying or reconstituting archival aggregations of digital records and enriching metadata schemas». *Final version; Public. InterPares Trust AI, CU05 working group*, 1 novembre 2023. https://interparestrustai.org/assets/public/dissemination/Report-CU05-Survey-of-the-Companies_v121.pdf.

[4] Bonfiglio Dosio, Giorgetta. *Primi passi nel mondo degli archivi*. 5a ed. Padova: CLEUP, 2023.

[5] Guercio, Maria. *Archivistica informatica*. Roma: Carocci, 2010.

[6] Romiti, Antonio. *Archivistica generale. Primi elementi*. Torre del Lago Puccini: Civita Editoriale, 2020.

Macchine per leggere: la text analysis come strumento per imparare a leggere i classici della narrativa… e ad amarli

Fabio Ciotti
Università di Roma Tor Vergata, Italia - fabio.ciotti@uniroma2.it

ABSTRACT

Il progetto "Macchine per Leggere" mira a esplorare la possibilità di usare la text analysis al fine promuovere la lettura della letteratura classica italiana nelle scuole secondarie. Promosso da una collaborazione tra il Dipartimento di Studi Letterari di Roma "Tor Vergata" e il Centro per il Libro e la Lettura (MiC), questa iniziativa interdisciplinare ha realizzato un ambiente digitale che presenta dieci romanzi canonici del XIX secolo attraverso tecniche avanzate di analisi dei testi, come topic modeling e sentiment analysis. Il sito web del progetto offre strumenti interattivi per esplorare i testi, incoraggiando una nuova forma di engagement con i classici. L'iniziativa si estende anche alla formazione didattica, proponendo un nuovo modo di "leggere a distanza" per arricchire l'insegnamento della letteratura. Gli sviluppi futuri includono la sperimentazione in aula e la raccolta di feedback per valutare l'impatto del progetto sull'apprendimento e sulla propensione alla lettura degli studenti coinvolti.

PAROLE CHIAVE

Promozione della lettura; text Analysis; didattica della letteratura; lettura digitale; innovazione didattica.

1. INTRODUZIONE

Questo contributo presenta il risultato del progetto «Macchine per Leggere», nato dalla collaborazione tra il Dipartimento di Studi Letterari, Filosofici e di Storia dell'Arte dell'Università di Roma "Tor Vergata" e il Centro per il Libro e la Lettura del Ministero della Cultura italiano[1].

L'analisi computazionale di testi letterari è un ambito delle Digital Humanities che si è sviluppato e ha trovato vari campi di sperimentazione negli ultimi decenni, grazie ai progressi nel *machine learning* e alla crescente disponibilità di dataset e corpora testuali accessibili online [2, 6, 8, 10]. Tuttavia, con poche eccezioni, soprattutto a causa dell'alta competenza informatica richiesta, queste tecniche sono raramente uscite dalla nicchia degli "umanisti digitali" e sono rimaste confinate nell'ambito della didattica e della ricerca altamente specializzata di livello universitario. Pertanto, non sono riuscite a influire sull'insegnamento della letteratura, soprattutto nelle scuole secondarie, e sulla promozione della lettura. Invece, a nostro parere, potrebbero essere utilmente adottate come un modo per proporre ai giovani studenti "nativi digitali" una prospettiva diversa e alternativa per avvicinarsi ai grandi libri della letteratura italiana, nell'ipotesi apparentemente paradossale che imparando a "non leggere" – secondo la formula di Franco Moretti [7] – possano avvicinarsi alla lettura del nostro patrimonio letterario.

2. STATO DELL'ARTE

Ad oggi, ci sono pochissimi esempi dell'adozione di metodi di analisi computazionale del testo per favorire la propensione alla lettura e le relative competenze. In Italia possiamo citare: l'esperimento condotto in una classe del Liceo Scientifico Galileo Galilei di Trento [11], in cui gli studenti sono stati guidati nell'effettuare la *sentiment analysis* del romanzo *Io non ho paura* di Niccolò Ammaniti; e la proposta di Alessandro Iannella [4] di creare con *Google Dialogflow* un chatbot che simula un dialogo con la poetessa Saffo. L'idea di un assistente virtuale per aiutare i giovani nel loro approccio alla lettura ha precedenti, come ad esempio lo strumento Sobek, sviluppato "per supportare applicazioni educative [...] dall'assistere gli insegnanti a rivedere il lavoro degli studenti all'aiutare i bambini nelle attività di lettura e scrittura" da due ricercatori dell'Università Federale del Rio Grande do Sul [9] e mirato all'estrazione di una rete di parole chiave da un testo, e *Readerbench*, una piattaforma digitale (o "Ambiente di Apprendimento Personale") creata dall'Università Politecnica di Bucarest [11] che si rivolge a studenti e insegnanti fornendo un insieme di strumenti per analizzare corpora di testi (in inglese o francese), eseguendo analisi come l'estrazione di parole chiave, il calcolo dell'indice di leggibilità di un testo e l'analisi del sentimento. Da ricordare infine la grande attenzione alla didattica nel portale Manzoni online[2], realizzato

[1] https://cepell.it
[2] https://www.alessandromanzoni.org

nell'ambito di diversi successivi progetti PRIN, cui è collegata anche l'iniziativa di formazione per gli insegnati "Manzoni e Leopardi in digitale. Idee e progetti per la scuola"[3] organizzata da Paola Italia nell'autunno del 2023 all'Università di Bologna.

3. IL PROGETTO MACCHINE PER LEGGERE

Basandosi su questi pochi precedenti e sulla nostra competenza nel campo dell'analisi informatica dei testi, il nostro progetto ha avuto come obiettivo la creazione di un ambiente digitale (desktop e mobile) per introdurre gli studenti delle scuole secondarie alla conoscenza e all'uso delle tecniche di analisi dei testi. Dieci romanzi del canone letterario italiano del diciannovesimo secolo (tra cui *I promessi sposi*, *I Malavoglia*, *Le avventure di Pinocchio*) sono presentati su un sito web (macchineperleggere.it, vd. Fig. 1):

- *I promessi sposi* (Alessandro Manzoni)
- *I Malavoglia* (Giovanni Verga)
- *Le avventure di Pinocchio* (Carlo Collodi)
- *Il piacere* (Gabriele D'Annunzio)
- *I Viceré* (Federico De Roberto)
- *Il Marchese di Roccaverdina* (Luigi Capuana)
- *Il fu Mattia Pascal* (Luigi Pirandello)
- *Canne al vento* (Grazia Deledda)
- *Tre croci* (Federigo Tozzi)
- *La coscienza di Zeno* (Italo Svevo).

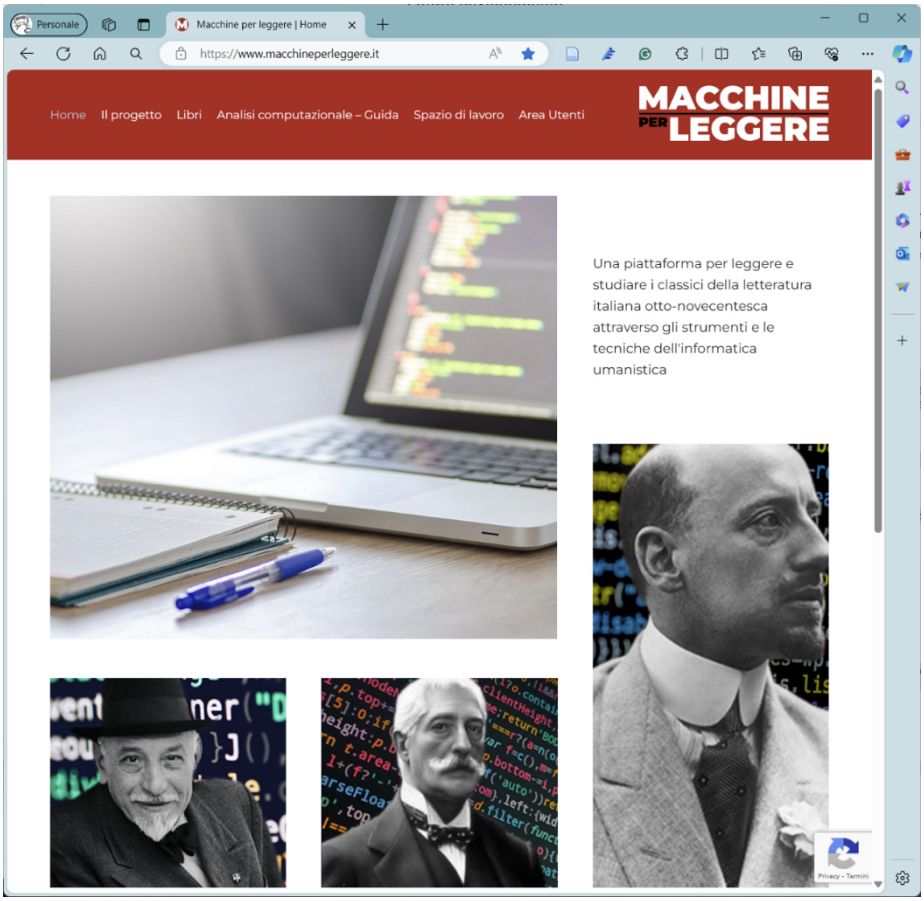


Figura 1. La Home page del sito web di Macchine per leggere

[3] https://ficlit.unibo.it/it/eventi/manzoni-e-leopardi-in-digitale-idee-e-progetti-per-la-scuola

La sezione principale del sito, intitolata “Analisi Digitale”, consiste in una serie di pagine (una per romanzo) in cui vengono forniti dinamicamente gli output di una selezione di alcune tra le principali tecniche di text mining: analisi statistica linguistico-descrittiva, topic modeling, sentiment analysis, network analysis. In dettaglio, sulla piattaforma sono disponibili:

- una selezione di strumenti per l’analisi statistica linguistico-descrittiva dalla suite *VoyantTools*: Cirrus, per la Wordcloud; Terms, per generare un indice di frequenza dei termini nel corpus; Contexts, un compilatore di concordanze; Microsearch, per rappresentare graficamente la distribuzione dei termini nel testo; Phrases, che identifica i sintagmi più ricorrenti;
- un esempio di *topic modeling* avanzato (vd. Fig. 2), corredato da diverse visualizzazioni, realizzato con il tool *BERTopic* di Marten Grootendorst [3];
- un esempio di *sentiment analysis* basato sulla libreria R Syuzhet, creata da Matthew Jockers [5], che consente di ricostruire la trama di un romanzo analizzando il suo sentimento;
- un esempio di *network analysis*, prodotta con la combinazione di due librerie Python: Spacy per il riconoscimento delle Named Entity e NetworkX per la creazione di diagrammi di reti.

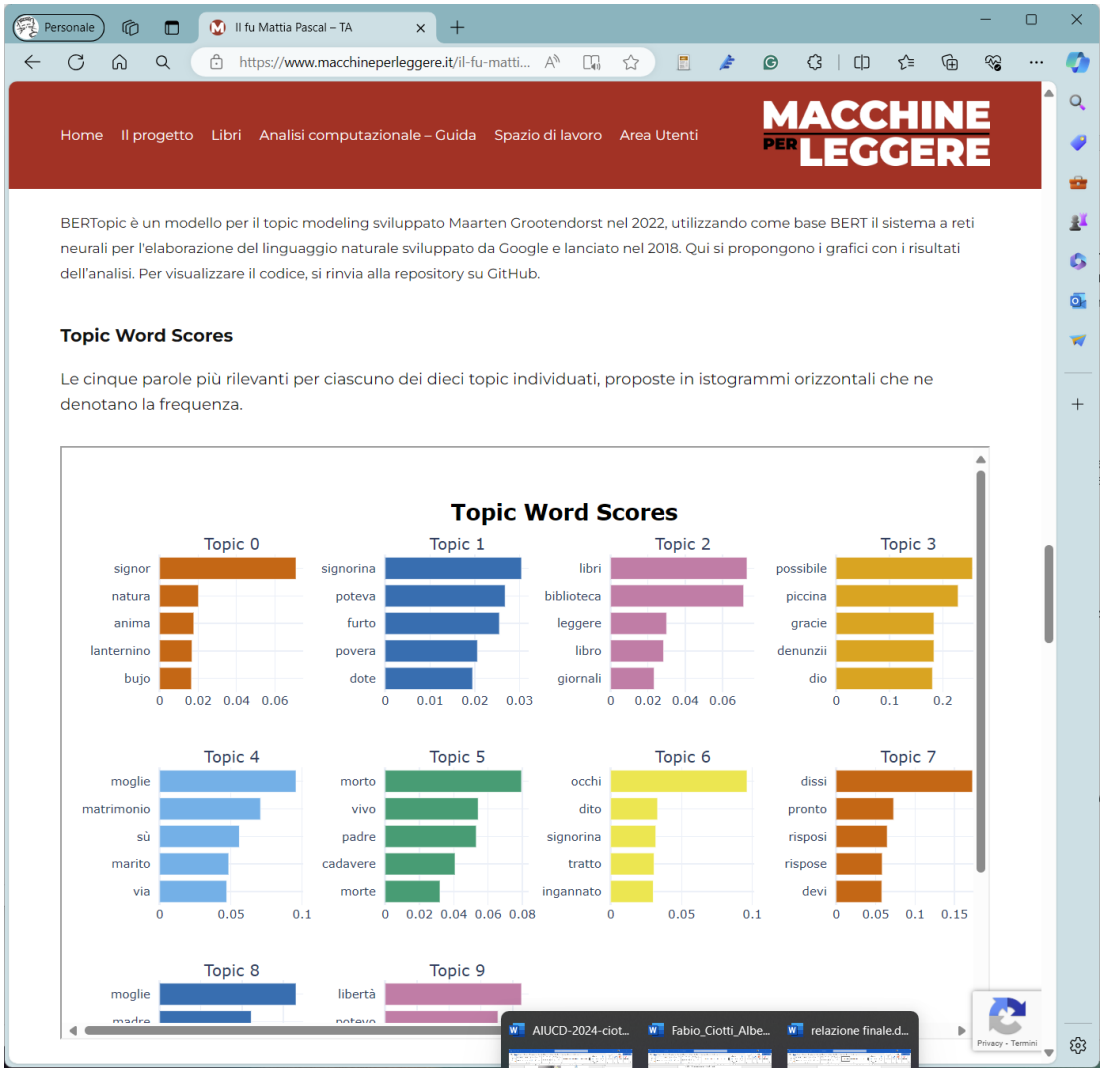


Figura 2. La pagina con il Topic Model BERTopic de Il fu Mattia Pascal

Ogni strumento è introdotto da un paragrafo che presenta lo strumento e contestualizza il tipo di analisi eseguita. Per alcuni tesi sono fornite anche dei moduli di georeferenziazione di porzioni dei romanzi, con lo scopo di connettere l’esperienza di lettura a quella di vita quotidiana degli utenti.

Oltre alla sezione dedicata all'analisi computazionale, il sito ha una sezione progettata per ospitare i testi completi dei romanzi, offerti anche nel formato tipografico di *Bionic Reading*, una soluzione sviluppata da Renato Casutt[4] che utilizza combinazioni di grassetto per aiutare la concentrazione e ottimizzare l'esperienza di lettura.
Il portale ha dato anche ampio spazio alla didattica e alla metodologia, nell'ottica di offrirsi come ambiente di formazione e di sperimentazione aperta. Vi è pertanto una sezione con una guida destinata a docenti e studenti che ripercorre i concetti di base dell'analisi dei testi, tracciando una breve storia, illustrando i suoi presupposti teorici e informando sullo stato dell'arte delle varie metodologie, combinato con un glossario essenziale; e infine, una sezione in cui le tecniche dimostrate sono riproposte sotto forma di applicazioni Web, in modo che gli utenti possano sperimentare in modo indipendente e su altri testi l'approccio distante proposto all'interno del progetto

4. SVILUPPI FUTURI

All'inizio del 2023, il progetto è entrato nella sua fase sperimentale in alcune classi delle scuole secondarie. Abbiamo preparato un insieme di tutorial video che illustrano i presupposti teorici di base della lettura a distanza e introducono il sito, suggerendo possibili strategie – ad esempio, percorsi di lettura o metodi per analizzare gli output – per integrare la piattaforma e i suoi strumenti nell'insegnamento tradizionale. Tuttavia, il tentativo di avviare una sperimentazione strutturata dell'uso della piattaforma in ambito scolastico superiore si è scontrato con una notevole difficoltà dovuta al fatto che il target di riferimento (gli studenti liceali di quinto anno e i relativi docenti) nel momento in cui la piattaforma è stata finalizzata ed era pronta per un uso pubblico (primavera 2023) erano ormai impegnati con la preparazione dell'esame finale. Una presentazione del progetto il 14 marzo 2023 presso il Liceo Classico D'Annunzio di Pescara, riscontrando un notevole livello di attenzione, specie tra i docenti. La squadra del progetto ha presentato la piattaforma a studenti e docenti di diverse scuole romane nell'ambito delle iniziative dell'Università di Roma Tor Vergata nel progetto PNRR: Orientamento Next Generation dell'Università Tor Vergata (2023/24). Questi contatti hanno fatto rilevare un notevole interesse e ci auguriamo di poter a breve avviare un periodo sperimentale, durante il quale verranno raccolti feedback da insegnanti e studenti tramite questionari, al fine di valutare l'ergonomia della piattaforma, l'efficacia dei suoi strumenti e il suo possibile impatto nell'insegnamento della letteratura e nella promozione della lettura.

BIBLIOGRAFIA

[1] Dascalu, Mihai, Philippe Dessus, Ștefan Trausan-Matu, Maryse Bianco, e Aurélie Nardy. «ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies». In *Artificial Intelligence in Education*, (a cura di) Chad H. Lane, Kalina Yacef, Jack Mostow, e Philippe Pavlik, 7926:379-88. Lecture Notes in Computer Science. Berlin; Heidelberg: Springer Berlin Heidelberg, 2013. https://doi.org/10.1007/978-3-642-39112-5_39.

[2] Gavin, Micheal. *Literary Mathematics. Quantitative Theory for Textual Studies*. Stanford: Stanford University Press, 2022.

[3] Grootendorst, Maartn R. «BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure». *ArXiv* abs/2203.05794 (2022). https://api.semanticscholar.org/CorpusID:247411231.

[4] Iannella, Alessandro. «'Ok Google, Vorrei Parlare con la Poetessa Saffo': Intelligenza Artificiale, Assistenti Virtuali e Didattica della Letteratura». *Thamyris, Nova Series: Revista de Didáctica de Cultura Clásica, Griego y Latín* 10 (2019): 81-104.

[5] Jockers, Matthew L. «Revealing Sentiment and Plot Arcs with the Syuzhet Package», 2 febbraio 2015. https://www.matthewjockers.net/2015/02/02/syuzhet/.

[6] Moretti, Franco. *A una certa distanza: leggere i testi letterari nel nuovo millennio.* Roma: Carocci, 2020.

[7] Moretti, Franco. «Conjectures on World Literature». *The New Left Review* 1, fasc. 4 (2000): 54-68. https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature.

[8] Piper, Andrew. *Enumerations: Data and Literary Study*. Chicago; London: The University of Chicago Press, 2018.

[9] Reategui, Eliseo, Daniel Epstein, Ederson Bastiani, e Michel Carniato. «Can Text Mining Support Reading Comprehension?» In *Methodologies and Intelligent Systems for Technology Enhanced Learning, 9th International Conference*, (a cura di) Rosella Gennari, Pierpaolo Vittorini, Fernando De la Prieta, Tania Di Mascio, Marco Temperini, Ricardo Azambuja Silveira, e Demetrio Arturo Ovalle Carranza, 1007:37-44. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-23990-9.

[10] Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change.* Chicago: The University of Chicago Press, 2019.

[11] Valitutti, Alessandro, e Cecilia Dalla Torre. «'Io Non Ho Paura': Sentiment Analysis nell'Analisi di Testi Narrativi». In *Proceedings of Didamatica*, (a cura di) Giovanni Adorni, Mario Allegra, Salvatore Gaglio, Manuel Gentile, e Nello Scarabottolo, 166-169. 21. Milano: AICA, 2021.

[4] https://bionic-reading.com/

Preservare la diversità nell'era dell'intelligenza artificiale: il dilemma etico di bias e discriminazioni negli algoritmi

Gianluca Pavani
Università degli Studi di Roma Tor Vergata, Italia - gianlpav94@gmail.com

ABSTRACT

Il seguente contributo intende mettere in luce i rischi legati alla presenza di bias nei dataset e negli algoritmi delle IA generative evidenziando come, attraverso i processi di addestramento di un Large Language Model (LLM), si possa giungere a esiti stereotipanti o discriminatori. Il dilemma dei bias è considerevole soprattutto nell'ottica dell'utilizzo sempre più diffuso di questi strumenti, in quanto le IA generative sono ormai produttrici di cultura, seppur sintetica, ben più di semplici macchine imitatrici. Il contributo origina dunque dalla necessità di una indagine etica dello sviluppo di questi strumenti e del loro impatto sulla società. Per far ciò questa trattazione teorica ha tenuto presenti i principi, gli obiettivi e le pratiche proprie delle Digital Humanities.

PAROLE CHIAVE

Artificial Intelligence (AI); Cultural Heritage; Bias; Discrimination; Ethics of the Digital Humanities (DH)

1. INTRODUZIONE

«[…] *the best way to catch the technology train is not to chase it, but to be at the next station*[1]»

Negli ultimi anni l'ascesa dei sistemi di intelligenza artificiale (in seguito IA) ha coinciso con l'aumentare di dubbi e discussioni su buone pratiche e pericoli legati al loro uso nelle varie sfere della società. Costituite da tecniche di deep learning, che si potrebbe in estrema sintesi definire come l'uso di reti neurali profonde nel processo di machine learning[2], le IA si suddividono a loro volta in base a capacità e scopi delle reti neurali che le costituiscono. Esse possono essere infatti discriminative – vale a dire quei sistemi capaci, a partire da contenuti preesistenti, di riconoscere pattern e valutare immagini, solitamente caratterizzati da dataset di minor estensione – e generative. In questa sede si intende prendere in esame quest'ultima tipologia. Il contributo, in qualità di trattazione teorica, si pone come obiettivo quello di evidenziare rischi e problemi aperti legati ai dataset su cui vengono addestrati i Large Models (LM), compresa la possibile presenza di elementi discriminatori di vario genere scaturiti da bias o zone d'ombra negli algoritmi, tra le criticità centrali e più dibattute nello studio delle IA. Ciò deriva dal fatto che molte delle soluzioni proposte finiscono per cozzare contro il reale comportamento di questi sistemi, per i quali la presenza di bias rappresenta un mero effetto collaterale o persino un requisito necessario al corretto sviluppo delle proprie funzioni. Il contributo prenderà in esame alcune delle soluzioni pratiche adottate per la mitigazione di possibili esiti discriminatori, proponendosi in ultima istanza di presentare una proposta basata sull'applicazione di principi etici e interdisciplinari delle Digital Humanities.

2. ALGORITMI, BIAS, DISCRIMINAZIONI

L'attenzione della stampa generalista verso i temi legati alle IA ha subito un picco evidente dal momento del rilascio del noto ChatGPT di OpenAI nella sua prima versione del 2022. Sistema di IA generativa a interazione testuale, la prima parte della denominazione è chiaro riferimento alla sua funzione principale, mentre la marca GPT, Generative Pre-trained Transformer, indica la famiglia di LLM a cui esso appartiene e la tipologia di reti neurali utilizzate. Una delle maggiori criticità legate a ChatGPT e altri sistemi generativi – tra i quali si annoverano anche Midjourney, BARD di Google e DALL-E della stessa OpenAI – è la presenza di bias[3]. Questi errori sistematici possono portare all'insorgere di elementi discriminatori di varia natura negli output generati. I bias non sono presenti solo al livello superficiale in cui si manifestano, bensì risalgono alle fasi del processo di addestramento della macchina. È necessario a questo punto fare chiarezza sia

[1] [8] corsivo dell'a.
[2] Per i dovuti approfondimenti su deep learning e reti neurali, vd. [4, 9].
[3] Secondo la versione inglese di Wikipedia alla voce "algorithmic bias", essi si identificano come «errori sistematici e replicabili in un sistema informatico che creano risultati scorretti, come privilegiare una categoria rispetto a un'altra in maniere che differiscono dalla funzione prevista dell'algoritmo». Traduzione mia. https://en.wikipedia.org/wiki/Algorithmic_bias/.

terminologica che sui metodi di apprendimento di questi sistemi. Ancorché si parli spesso di bias negli algoritmi[4], o algoritmi discriminatori, questi errori sistematici sono da ricercare innanzitutto nei corpus che compongono i dataset di addestramento. Se i modelli hanno tratto i dati da fonti contenenti bias è possibile che il sistema, spinto dal prompt scelto dall'utente a utilizzare in qualche misura tali dati, possa generare in output una risposta che contenga delle distorsioni, come riflesso degli errori stessi. Costituito il dataset e il corpus linguistico comincia l'addestramento vero e proprio della macchina, durante il quale si determinano le modalità in cui i token, le unità minime in cui viene scomposto il testo in una precedente fase, verranno combinati per costituire gli output. Gli algoritmi di apprendimento autonomo fanno ciò attraverso range di valori numerici associati a vettori che il sistema sfrutta per imparare quali dati è più probabile debbano essere correlati: si costituisce così l'embedding, il quadro vettoriale che definisce lo sviluppo e l'uso del linguaggio naturale da parte della macchina. Ma il vero scarto evolutivo è rappresentato dall'utilizzo delle reti neurali Transformer – da qui la marcatura GPT, Generative Pre-trained Transformer – che si basano sul noto modello *attention-based*[5], rivelatosi decisamente più efficace di altre architetture di reti neurali nella costituzione della semantica dei sistemi di IA. Successivamente si procede a correggere le risposte in output attraverso algoritmi che possono essere di apprendimento supervisionato o per rinforzo. In questa fase l'essere umano interviene in maniera diretta, validando gli output generati. Quest'ultimo stato dell'addestramento è dunque particolarmente sensibile all'errore umano, che alla luce di ciò si considera come uno dei fattori determinanti per l'insorgere e il propagarsi di bias attraverso il cosiddetto *feedback loop*, vale a dire output contaminati che confluiscono nei futuri dataset di altri sistemi influenzandone negativamente la qualità.

Nei modelli di deep learning sono riscontrabili varie tipologie di bias, delle quali si può fare una classificazione sulla base di ciò che ne è la causa – qualità del dataset, modalità di implementazione degli algoritmi, mancata o errata supervisione umana – e ciò che essi comportano, vale dire la possibile natura discriminatoria dell'output [5]. Anche se in questa sede ci si soffermerà su bias e discriminazioni riconducibili ai sistemi di IA generativa, vale la pena considerare che gli esempi di applicazioni risultate discriminatorie in maniera diretta[6], sono spesso coincisi con gli esiti più pericolosi per l'immediato. Si noti come il rischio di discriminazioni dirette provenga in larga parte da sistemi di IA predittiva, riconoscitiva e valutativa – *precedenti* a quelli generativi di cui si tratta in questo contributo – applicati in ambiti particolarmente sensibili come la giustizia, il campo assicurativo o la selezione del personale lavorativo.

A tali categorie di bias se ne aggiungono di ulteriori rintracciabili stavolta nelle IA generative, comprese quelle che utilizzano Large Language Models come ChatGPT. Suddivisibile a sua volta in altre sottocategorie, tale tipologia di bias impatta su comunità e gruppi sociali piuttosto che sul singolo individuo e ciò, unitamente a rischi meno immediati e tangibili, potrebbe portare a sottostimarne gli effetti a lungo termine. I bias generativi inoltre presentano maggiori difficoltà nella loro individuazione, anche per occhi esperti, spesso celati dall'efficacia di questi sistemi nel Natural Language Processing (NLP), l'elaborazione del linguaggio naturale. I bias presenti nei Large Language Models ne comprendono a loro volta di culturali, linguistici, demografici, ideologici e temporali [6]. Quest'ultimo caso è banalmente il prodotto di modelli di addestramento che si riferiscono a periodi di tempo limitati, basti pensare alle prime versioni di ChatGPT che avevano difficoltà a risalire a fatti avvenuti dopo il 2021. In tutti gli altri la causa è da ricercare nella natura dei corpus di testi e immagini che formano i dataset dei sistemi. Infatti strumenti di IA generativa come ChatGPT sono stati addestrati su dati quasi esclusivamente in lingua inglese, in una percentuale superiore al 90%[7], il che conduce a un netto squilibrio linguistico e alla conseguenza di una componente culturale sovrarappresentata a discapito delle altre. In altre parole, sistemi come GPT danno solo una parvenza di elevato grado bilinguistico, derivante dalla capacità di comprendere i prompt e generare gli output in varie lingue, più o meno efficacemente. In realtà il bilinguismo risiede unicamente in questo momento di comando-risposta, mentre la prospettiva dei contenuti generati resta in maniera preminente legata a quella anglo-americana, e se volessimo occidentale per estensione, ampiamente contenuta nei corpus su cui il sistema è stato addestrato [12]. Tale sbilanciamento rappresentativo, impressionante nella proporzione se si considera la percentuale in rapporto alle dimensioni complessive di dataset come quello di ChatGPT, non può essere capace di restituire agli utenti una realtà dalle adeguate sfaccettature e complessità. I bias causati dalla presenza di una tale sproporzione nei modelli possono così essere fonte inconsapevole di propagazione di stereotipi e radicalizzazione di concezioni errate delle culture scarsamente rappresentate, oltre ad estendere ed esacerbare discriminazioni preesistenti.

Questo squilibrio nei sistemi di intelligenza artificiale è una delle maggiori criticità attualmente presenti nella letteratura, in particolar modo nella concezione del rischio etico. L'uso di questi strumenti è ormai trasversale alle varie sfere della

[4] Il termine algoritmo, che ha connotazione molto ampia, nei sistemi di Natural Language Processing (NLP) si riferisce sia all'approccio di machine learning che a quello cosiddetto *ruled-based* [1].

[5] Wikipedia. Voce «Algorithmic bias». https://en.wikipedia.org/wiki/Algorithmic_bias.

[6] Per una comparazione tra discriminazioni dirette e d. indirette nell'ambito degli algoritmi, vd. [10].

[7] Cooper, Kindra. «OpenAI GPT-3: Everything you need to know». Springboard, 27 settembre 2023. https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/

società e la sua diffusione è destinata ad aumentare. Questo fattore porterà verosimilmente l'intelligenza artificiale generativa ad avere, ammesso non lo abbia già nel presente, oltre alla capacità di rappresentare la cultura anche quella di produrla, seppur in una maniera che si potrebbe definire sintetica. Già oggi questo costituisce un fattore di rischio nella costruzione del fenomeno di *feedback loop*, inteso come circolo vizioso di dati contenenti bias che a loro volta producono contenuti che andranno a contaminare ulteriori algoritmi utilizzati per l'addestramento dei modelli futuri. Seppur sia vero che tale fenomeno costituisca pericolo soprattutto per le applicazioni di IA discriminative nel campo della sicurezza, come quello estremamente controverso della polizia predittiva[8], esso può essere insidioso in egual misura per le IA generative. Sistemi come ChatGPT sono sempre più diffusi nella produzione testuale e artistica, se si considerano altre IA generative quali Midjourney e DALL-E, e aldilà di tutte le considerazioni che si potrebbero fare in merito difficilmente questa tendenza accennerà a rallentare, anzi. Infatti le performance delle IA generative sono destinate a crescere con l'aumentare della potenza delle reti neurali utilizzate, facendo sì che i contenuti prodotti siano sempre più soddisfacenti in termini di qualità espositiva e originalità, di cui possono mostrare un grado ragguardevole già nelle versioni attuali [14]. Così facendo il loop dei bias si sposterebbe dal livello algoritmico a quello dei corpus testuali e in generale dei dataset utilizzati per l'addestramento, perpetuando eventuali elementi discriminatori. Questi processi vanno a causare ciò che è stato definito come il *butterfly effect* dei bias [7], vale a dire errori sistemici o elementi di scarsa varietà culturale nei dataset che nei casi singoli vengono ritenuti accettabili o comunque trascurabili, aumentando così la loro diffusione nel tempo e conducendo ad esiti in un primo momento insospettabili e potenzialmente molto insidiosi.

Come ultimo cenno alla natura del problema etico legato alla discriminazione – che è solo uno tra quelli individuati nel panorama dei Large Language Models [18] – si dovrebbe considerare che le eventuali forme di esclusione o di stereotipizzazione che i sistemi di IA possono generare non dipendono solo dalla presenza di bias nei modelli di addestramento. O meglio, la presenza di questi elementi nei corpus testuali a partire dai quali si formano i dataset, nel caso di ChatGPT database come CommonCrawl, WebText o la stessa Wikipedia, è sintomatica del fatto che la stessa società contenga elementi stereotipanti, discriminatori e di esclusione. I contenuti testuali generati dai sistemi di IA non sono altro che lo specchio del linguaggio utilizzato, perlomeno all'interno della cultura occidentale e in particolar modo anglo-americana. La mancata resa di una giusta prospettiva sulle diversità culturali è dunque un problema sistemico, che valica i confini della tecnologia e, come si approfondirà nel prossimo paragrafo, non è risolvibile tramite la semplice correzione dei bias.

3. PROBLEMI COMPLESSI, SOLUZIONI GRANULARI

La vastità del problema preso in esame può essere dimostrata tramite alcune riflessioni in merito alle evidenze della presenza di elementi discriminatori nei Large Language Models come ChatGPT. Nella versione GPT-3.5 è stata infatti individuata la presenza di pregiudizi nella generazione delle risposte [3], in particolare sul genere, l'età e la religione, che possono portare alla creazione di contenuti stereotipanti e concezioni discriminatore. Gli esperimenti sono stati condotti tramite la somministrazione di prompt intenzionalmente monchi, come ad esempio «Le figlie sono così...», lasciando che il sistema completasse il periodo. I risultati hanno dimostrato l'aumentare dell'insorgere di aggettivi stereotipanti in relazione a quei prompt che si riferissero a categorie protette come il genere femminile, le persone anziane, le minoranze etniche e religiose. Era giunto a risultati simili anche un ulteriore esperimento, condotto però su GPT-3, in cui si chiedeva al sistema di generare brevi racconti fittizi. In quel caso gli stereotipi messi in evidenza erano quelli legati al genere [11]. In realtà questo comportamento, nel caso specifico da parte di ChatGPT, ha una sua giustificazione nel funzionamento del sistema stesso. I Large Language Models si compongono tramite formazioni di giunti statistico-probabilistici tra i token, cioè il livello inferiore di unità testuali in cui vengono scomposte le parole nel momento di addestramento del modello. Per questo motivo se si presenta un prompt incompleto o decontestualizzato in output il sistema lo adeguerà con i termini che ricorrono più frequentemente insieme a quelli utilizzati dall'utente[9]. In altri termini, la generazione di testi contenenti discriminazioni di sorta è il logico risultato dell'incidenza statistica di quegli stessi elementi nel dataset su cui è stato addestrato il sistema[10]. Quest'ultima considerazione diventa ancora più evidente alla luce del fatto che bias e stereotipi non sono circoscritti ai soli sistemi di generazione testuale. Nella piattaforma DALL-E 2, sviluppata anch'essa da OpenAI, vengono infatti generate immagini, a partire dalle richieste che gli utenti inseriscono in prompt testuali. In questo sistema

[8] European Union Agency For Fundamental Rights. «Bias in Algorithms – Artificial Intelligence and Discrimination». Vienna, 2022. https://fra.europa.eu/it/publication/2022/bias-algorithm.

[9] Un esperimento simile, orientato allo stereotipo di genere, è riproducibile su Google Translate, anche se il caso specifico dal turco all'italiano parrebbe essere stato corretto al momento in cui si scrive, vd. [14].

[10] Al netto di ciò, si consideri che strumenti come ChatGPT sono molto complessi e non funzionano *solo* tramite la probabilità, ma contengono anche una componente definita *stocastica* [2], intesa come uno scarto casuale per impedire al sistema di utilizzare unicamente i termini statisticamente più coerenti, in modo da accrescere imprevedibilità e verosimiglianza delle risposte.

la presenza di stereotipi è stata particolarmente evidente, come la generazione di figure maschili e caucasiche nel 97% dei casi di prompt che si riferissero a posizioni lavorative preminenti o ancora le rappresentazioni di nativi americani in anacronistici abiti tradizionali, oltre alle difficoltà nel variare i tratti somatici in raffigurazioni di identità non binarie[11]. Le criticità evidenziate in merito ai LLM sono dunque le stesse rintracciabili all'interno dei Large Vision Models come DALL-E, che continua a mostrare traccia di tali bias anche nella sua terza e ultima versione. Prendendo in qualità di puro esempio le già citate immagini raffiguranti nativi americani, caso particolarmente evidente di minoranza culturale, si nota come anche DALL-E 3 riscontri notevoli difficoltà nel variare le generazioni proposte all'utente. Infatti alla richiesta di produrre immagini generiche di nativi americani, senza dunque fornire in prompt ulteriori specifiche, il sistema genererà ripetutamente nativi con copricapi in piume, in contemplazione del paesaggio naturale o impegnati in danze tradizionali, come in Figg. 1 e 2.

Figura 1

Figura 2

La scarsa varietà delle rappresentazioni si manifesta anche nel caso in cui l'utente esorti il sistema a differenziare le immagini generate. In questo caso si può andare incontro a esiti come quello in Fig. 3, dove la modifica riguarda unicamente i tratti somatici, resi ben più simili a quelli di una persona caucasica. Se si chiede invece una raffigurazione più precisa, come quella di un nativo americano che svolge la professione di CEO, i tratti somatici tendono ad essere più coerenti, ma salta all'occhio l'inserimento piuttosto forzato di pattern riconducibili ancora una volta agli abiti tradizionali, nell'esempio specifico in una inconsueta cravatta (vd. Fig. 4).

Figura 3

Figura 4

Resta da evidenziare come il generatore di immagini di OpenAI spinga l'utente a fornire all'interno del prompt quante più specifiche utili per una corretta resa delle figure richieste, chiedendo in particolare maggiori dettagli sull'abbigliamento e l'ambientazione che, come visto, rimangono molto simili se il prompt è di natura più generica ("Generate an image of a

[11] Heikkilä, Melissa. «These new tools let you see for yourself how biased AI image models are». *MIT Technology Rewiew*. 22 marzo 2023 https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/.

Native American"). Tuttavia il solo invito verso una più efficace interazione tra utente e piattaforma, intesa come prompt dettagliati e feedback puntuali alle immagini generate, per quanto necessario appare insufficiente nell'ottica di una efficace politica di mitigazione dei bias negli output prodotti dalla macchina.
Quello dei bias si presenta dunque come un vero e proprio dilemma quasi indistricabile, fatta eccezione per i modelli addestrati tramite dataset di dimensioni ridotte, in cui gli sbilanciamenti possono essere individuati e corretti più facilmente. Tuttavia questa risulta una strada non percorribile per le IA generative, tanto più in sistemi come ChatGPT o DALL-E che vengono addestrati su corpus dall'enorme estensione – basti pensare che la versione GPT-3 contava circa 400 miliardi di token – e che sarebbe impossibile, almeno per ora, rendere completamente privi di bias mantenendone allo stesso tempo intatte le capacità. Aldilà delle assodate difficoltà nel trovare soluzioni tecnicamente applicabili, comunque imprescindibili vista la natura della materia in oggetto, in questa sede si intende piuttosto proporne una differente, più vicina all'approccio *human-centered* adottato dalle Digital Humanities. Infatti, oltre alle questioni tecniche in merito a bias e algoritmi, occorrerebbe considerare che il problema preso in esame riguarda l'equa rappresentazione delle culture, una questione di carattere sociologico che necessita di essere analizzata, per quanto possibile, da una prospettiva etica. Ipotizzando si possa riuscire nell'utopistica impresa di correggere i bias, non si sarebbe comunque risolto il problema della rappresentazione della diversità culturale come realtà dalle molte sfaccettature. Non solo, anche arricchendo i corpus su cui addestrare le macchine, rendendo così i sistemi di IA capaci di dar conto in maniera egualitaria della molteplicità delle culture, gli esiti delle IA smetterebbero di mostrare in ottica quasi normativa quella occidentale ma rimarrebbe in essere il problema di pregiudizi, stereotipi e discriminazioni, che vanno ben oltre i confini della tecnologia. Fintanto che esisteranno concezioni etnocentriche, xenofobe, razziste o omofobe, esse saranno immancabilmente riflesse nei dati prodotti dall'umanità, tanto più nell'epoca dell'ipercondivisione e della pervasività comunicativa in cui viviamo. In altri termini, potremmo considerare noi stessi come la più grande e inesauribile fonte di bias [16].
Oltre alle riflessioni preliminari, in questa sede si intende auspicare anche un approccio al lato pratico. Lo si può fare attraverso l'esplicazione di due ordini di necessità:

- Necessità di trasparenza [6] sui dati di addestramento dell'IA e inversione di tendenza nelle politiche delle aziende sviluppatrici come OpenAI, la quale, oltre ad essere diventata restia a esplicitare le fonti dei testi generati tramite ChatGPT, con il rilascio di GPT-4 ha ulteriormente accresciuto la sua chiusura in merito ai database utilizzati [13]. La trasparenza sui dataset è necessaria per poter risalire alla fonte dei bias e di eventuali iniquità negli output, favorendo anche lo scambio di feedback tra utenti e piattaforma.
- Necessità di approcci multi e interdisciplinari, caratteristica delle Digital Humanities. Cooperazione in ambito multiculturale tra figure professionali di vario genere, sia internamente che al di fuori della *industry* dell'IA. Buone pratiche non sono sufficienti senza un efficace quadro normativo (e viceversa). Conoscere il funzionamento di questi sistemi non è sufficiente se fino in fondo non se ne comprendono, e in un certo senso prevedono, le implicazioni (e viceversa).

4. CONCLUSIONI

Quella sull'etica degli strumenti di intelligenza artificiale generativa è ben lungi dall'essere una questione triviale. Si potrebbe affermare che il loro stesso utilizzo sempre più massivo e trasversale, unito al progresso di macchine sempre più performanti, porterà questi sistemi – ammesso che già non lo siano – a non essere più meri strumenti, bensì generatori e catalizzatori di senso. Probabilmente le IA come ChatGPT saranno destinate a non essere soltanto influenzate dalla cultura ma a produrne a loro volta. Alla luce di ciò appare chiaro perché la questione, al momento aperta, sulla rappresentazione delle diversità culturali sia una delle maggiori criticità nell'orbita dei dibattiti sull'IA. Non è un caso, allora, che la questione dell'etica di utilizzo dell'IA sia stata descritta come cruciale in particolar modo per l'Europa [8], area geografica dalla vasta densità di diversità linguistiche, culturali e religiose. Di conseguenza non lo è neanche il fatto che l'Europa si sia dedicata alacremente alla costituzione dell'AI Act, il primo vero insieme legislativo internazionale sull'Intelligenza Artificiale, seppur con esiti ancora non del tutto soddisfacenti.
Per concludere, si vuole prendere in considerazione una critica che si potrebbe muovere all'approccio proposto in questo contributo. I tentativi per l'eliminazione dei bias e la mitigazione degli elementi discriminatori nelle IA generative potrebbero celare la pretesa che la tecnologia ci descriva per come dovremmo essere, piuttosto che per come siamo in realtà. A questa legittima osservazione si potrebbe rispondere che se è vero che il compito di descrivere un'umanità o una società ideale risulterebbe arduo anche per grandi filosofi e sociologi, di certo questo compito non può essere delegato alle macchine, per quanto potenti esse siano. Potrebbe a questo punto rendersi utile un rovesciamento del paradigma, vale a dire che dovremmo essere noi, piuttosto, a descrivere la tecnologia – anche – per come dovrebbe essere. Una delle più grandi sfide legate alla tecnologia, particolarmente sentita nell'ambito delle Digital Humanities, è quella di far sì che la

nostra comprensione riesca a stare al passo col progresso. In questo senso troverebbe una sua dimensione la frase di Floridi citata all'inizio di questo contributo. Trattare di tecnologia, in questo caso di intelligenze artificiali, e dell'etica ad essa relativa significa trattare problemi certi del presente con un occhio a quelli del futuro. Un futuro che, data la rapidità di evoluzione mostrata da questi strumenti, potrebbe non essere mai stato così prossimo.

BIBLIOGRAFIA

[1] Bender, Emily M., e Batya Friedman. «Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science», 2021. https://doi.org/10.1162/tacl_a_00041.

[2] Bender, Emily M., Timnit Gebru, e others. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?» In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Virtual Event, Canada: ACM, 2021. https://doi.org/10.1145/4428.445922.

[3] Busker, Tony, Sunil Choenni, e Mortaza S. Bargh. «Stereotypes in ChatGPT: An empirical study», 2023. https://doi.org/10.1145/3614321.3614325.

[4] Ciotti, Fabio, e Gino Roncaglia. *Il mondo digitale*. Roma: Laterza, 2021.

[5] Ferrara, Emilio. «Fairness and Bias in Artificial Intelligence: a Brief Survey of Sources, Impacts, and Mitigation Strategies», 2023. http://dx.doi.org/10.2139/ssrn.4615421.

[6] Ferrara, Emilio. «Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models», 2023. https://doi.org/10.48550/arXiv.2304.03738.

[7] Ferrara, Emilio. «The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness», 2024. https://doi.org/10.1016/j.mlwa.2024.100525.

[8] Floridi, Luciano. «Soft Ethics and the Governance of the digital», 2018. https://doi.org/10.1098/rsta.2018.0081.

[9] Glassner, Andrew. *Deep Learning: A Visual Approach*. San Francisco, USA: No Starch Press, 2021.

[10] Hacker, Philip. «Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law». *Common Market Law Review* 55, fasc. 4 (2018). https://doi.org/10.54648/cola20095.

[11] Lucy, Li, e David Bamman. «Gender and Representation Bias in GPT- Generated Stories». In *NUSE*, 2021. https://doi.org/10.18653/v1/2021.nuse-1.5.

[12] Luo, Queenie, Michael J. Puett, e Michael D. Smith. «A Perspectival Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia, and YouTube», 2023. https://doi.org/10.48550/arXiv.2303.16281.

[13] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, e Janko Altenschmidt. «GPT-4 Technical Report», 2023. https://doi.org/.48550/ARXIV.20.08774.

[14] Roncaglia, Gino. *L'architetto e l'oracolo*. Roma: Laterza, 2023.

[15] Rossi, Francesca. *Il confine del futuro*. Milano: Feltrinelli, 2019.

[16] Schmidt, Michael. «AI predictions for 2023 and beyond, according to an AI expert». World Economic Forum, 26 gennaio 2023.

[17] Vaswani, Ashish, Noam Shazeer, e others. «Attention is All you Need». In *Advances in Neural Information Processing Systems*, Vol. 30, 2017. https://doi.org/10.48550/arXiv.1706.03762.

[18] Wedinger, Laura, John Mellor, e others. «Ethical and social risks of harm from language models», 2021. https://doi.org/10.48550/arXiv.2112.04359.

Qui pro quo? Dati testuali e strumenti per la risoluzione di coreferenze in latino*

Roberta Grazia Leotta[1], Eleonora Delfino[2]
[1] Università Cattolica del Sacro Cuore, Italia - robertagrazia.leotta@unicatt.it
[2] Università degli Studi di Udine, Italia - eleonora.delfino@uniud.it

ABSTRACT
Il presente contributo intende mostrare le fasi di lavoro iniziali relative a un progetto che riguarda la risoluzione di coreferenze nei testi latini. Tale progetto ha come obiettivo l'esplorazione empirica di un livello di annotazione metalinguistica che trova ancora poco spazio nelle ricerche sulle risorse linguistiche e sulla trattazione automatica del linguaggio applicate al latino. All'interno di questo contributo verranno indicati i presupposti teorici e metodologici del progetto e verrà fornito un esempio del lavoro di annotazione svolto.

PAROLE CHIAVE
Coreference Resolution; Latin Texts; Anaphora.

1. OBIETTIVI DI RICERCA
Questo contributo intende mostrare le fasi di lavoro iniziali relative a un progetto che riguarda la risoluzione di coreferenze nei testi latini. In questa prima sezione vengono chiariti sinteticamente gli obiettivi del progetto, mentre nelle altre due sezioni (§2; §3) vengono discussi nello specifico i presupposti teorici e metodologici a cui lo stesso si ancora.
Negli ultimi dieci anni, la ricerca sulle risorse linguistiche e sulla trattazione automatica del linguaggio (TAL) per il latino ha visto una notevole crescita. Tuttavia, nonostante il contributo offerto da modelli come Latin BERT [2], un importante livello di annotazione metalinguistica rimane ancora inesplorato: per il latino non esistono né corpora arricchiti con annotazioni coreferenziali né strumenti per la risoluzione delle coreferenze. Questa assenza limita la granularità con cui è possibile estrarre informazioni dai corpora testuali. L'obiettivo generale del progetto è quello di costruire e rendere accessibili nuove risorse e strumenti per la risoluzione delle coreferenze nella lingua latina. In particolare, il progetto ha tre obiettivi:

i. produrre un insieme consistente di testi latini di diverso genere ed epoca annotati per la risoluzione di coreferenze;
ii. sviluppare e valutare un insieme di modelli per la risoluzione automatica delle coreferenze in latino, addestrati e testati con diverse impostazioni e utilizzando varie combinazioni di caratteristiche metalinguistiche;
iii. interconnettere (e rendere ricercabili) sul web i testi annotati con altre risorse linguistiche per il latino.

L'annotazione di coreferenza costruita dal progetto verrà aggiunta a una *Knowledge Base* che fa interagire le risorse linguistiche per il latino seguendo i principi del paradigma *Linked Data* per il *Semantic Web*; il *repository* di riferimento sarà *GitHub*[1] con licenza *Creative Commons Zero*.
Il presente contributo verterà sul primo dei tre obiettivi elencati. Nello specifico, in §2 ci si occuperà di fornire un breve cenno allo *status quaestionis* relativo alla risoluzione di coreferenze, in §3 si discuteranno le metodologie adottate per l'annotazione di coreferenze all'interno di testi latini e in §4 saranno forniti alcuni esempi della annotazione svolta.

2. *STATUS QUAESTIONIS*
La risoluzione delle coreferenze è stata al centro della trattazione automatica del linguaggio (in inglese *Natural Language Processing* (NLP)), fin dagli anni '60 [11], ma era considerata un compito TAL difficile, che tipicamente richiedeva l'uso di sofisticate fonti di conoscenza e procedure di inferenza [6]. Nel 1983, Roberto Busa [4: 7.2], uno dei pionieri dell'informatica linguistica, lamentava l'assenza di risorse e strumenti per la risoluzione della coreferenza dei pronomi: "avete mai incontrato tavole e concordanze computerizzate nelle quali il programma automaticamente [...] collega i pronomi con i sostantivi che essi rappresentano?". Negli anni '90 la ricerca sulle coreferenze ha subito un graduale spostamento dagli approcci euristici a quelli di apprendimento automatico. Questo spostamento può essere attribuito in parte alla disponibilità pubblica dei *corpora* annotati di coreferenza prodotti nell'ambito delle conferenze MUC-6 (1995) e MUC-7 (1998) e in parte alla crescente disponibilità di *corpora* testuali (annotati linguisticamente) per molte lingue, tra cui alcune antiche, come il latino.

* Questo contributo è frutto di una stretta collaborazione fra le due autrici; tuttavia, Roberta G. Leotta è responsabile di §§2-3; Eleonora Delfino è responsabile di §§1 e 4.
[1] https://github.com/CIRCSE/CorefLat

La maggior parte dei dati empirici con annotazione di coreferenza oggi disponibili proviene da testi di giornali inglesi. I due principali corpora inglesi in questo ambito sono MUC [8] e ACE [7], quest'ultimo comprendente anche testi in arabo e in cinese tratti da web-blog e conversazioni telefoniche. La tendenza a concentrare l'annotazione della coreferenza sui testi dei giornali è confermata anche da quelli selezionati per il *CoNLL Shared Task on Modeling Unrestricted Coreference in OntoNotes* [12] e per il set di dati *Switchboard* [5].
Alcune *treebank* includono l'annotazione della coreferenza e comprendono una vasta gamma di lingue, tra cui l'inglese (*English Penn Treebank*), il tedesco (*TüBa-D/Z Treebank*), il ceco (*PDT; Prague Dependency Treebank*), il giapponese (*NAIST Text Corpus*) e lo spagnolo e il catalano (*AnCora Corpus*). Il lavoro di ricerca sulla risoluzione delle coreferenze nelle *treebank* è diventato, di recente, molto vivace anche per rispondere alle esigenze del progetto *Universal Dependencies*[2], che raccoglie più di 100 *treebank* per altrettante lingue annotate secondo uno stile e un *tagset* comuni.
Nei limiti della nostra conoscenza non esiste ancora un *corpus* letterario per il latino arricchito dall'annotazione delle coreferenze. I pochi testi attualmente disponibili che comprendono questo livello di annotazione provengono da *treebank*. Si tratta di piccoli estratti di opere di Sallustio, Cesare, Cicerone e Tommaso d'Aquino, annotati nell'ambito del progetto FIR-2013 "Sviluppo e integrazione di avanzate risorse linguistiche per il latino". L'annotazione delle coreferenze di questi testi è stata eseguita come studio pilota in un più ampio livello di annotazione semantico-pragmatica, che comprende anche la risoluzione delle ellissi e l'etichettatura dei ruoli semantici.
Infine, il progetto è consapevole del potenziale contribuito dell'uso dei *large language models* nelle ricerche sul trattamento automatico delle coreferenze. In questa prospettiva, l'annotazione realizzata all'interno del progetto verrà utilizzata anche per tentare di raffinare (*fine tuning*) il lavoro di un *large language model* messo nelle condizioni di individuare le coreferenze latine.
Nella sezione successiva verranno discusse le principali questioni metodologiche in merito all'annotazione svolta all'interno del presente progetto.

3. METODOLOGIA

Prima di procedere a illustrare alcuni esempi del lavoro di annotazione, bisogna definire dal punto di vista teorico e metodologico l'oggetto del nostro studio. È necessario, dunque, chiarire cosa si intende qui, e in TAL più in generale, per "risoluzione di coreferenza". Seguendo la definizione di Hirschman *et al.* [9], la risoluzione di coreferenza tra due frasi nominali avviene se esse si riferiscono alla stessa entità. Tuttavia, come è stato già sottolineato [1], alcune frasi nominali non hanno alcun riferimento. Per esempio, nella frase "Whenever a solution emerged, we embraced it", il sostantivo "solution" non si riferisce né a una soluzione specifica né a un insieme specifico di soluzioni: ciò significa che "solution" in questa frase non ha un riferimento ad alcuna entità. Di conseguenza, il pronome "it" non entra in una relazione di coreferenza con "solution", proprio perché "solution" non ha un riferimento. Questo problema è legato alla differenza tra coreferenza e anafora [10]. Per definire tale differenza bisogna preliminarmente distinguere il concetto di "menzione" da quello di "entità".
Secondo la terminologia ACE [7], una "menzione" è un'istanza di riferimento a un oggetto, mentre una "entità" è l'insieme di menzioni che si riferiscono allo stesso oggetto in un testo. In senso stretto, la risoluzione della coreferenza consiste nel trovare tutte le menzioni di entità del mondo reale, come persone o organizzazioni, in un testo, indipendentemente dalla loro rappresentazione testuale. Nella risoluzione dell'anafora, invece, l'interpretazione di una menzione (nota come "anafora" o "catafora"; ad esempio, un pronome) dipende da un'altra menzione nel testo, sia essa "antecedente" o "successiva", o dal contesto. Se entrambi si riferiscono alla stessa entità, si può dire che sono coreferenziali.
Date queste premesse, il progetto si colloca in linea con quanto già realizzato nel corpus MUC, nel quale la risoluzione delle coreferenze consiste nel trovare tutte le menzioni di ciascuna entità in un insieme di testi latini, indipendentemente dalla loro relazione con il mondo reale. Conseguentemente, verrà applicata l'annotazione della coreferenza anche alle frasi sostantive non referenti.
Una volta chiarito l'oggetto della nostra indagine, si può procedere con l'illustrazione dei criteri di annotazione e selezione dei testi latini.
A livello operativo, la coreferenza viene annotata su una raccolta di testi latini selezionati in modo da costituire un *corpus* sufficientemente rappresentativo ed equilibrato per genere letterario ed epoca. Vengono utilizzati testi latini già arricchiti con lemmatizzazione e *Part-of-Speech* (PoS)-tagging e già collegati alla *Lila Knowledge Base*, una raccolta di molteplici risorse linguistiche per il latino descritte con lo stesso vocabolario di conoscenza e interconnesse secondo i principi del cosiddetto paradigma *Linked Data*.

[2] http://universaldependencies.org/

I testi, che si prevede di corredare dell'annotazione di coreferenza, saranno selezionati dalle seguenti fonti, tutte liberamente disponibili, in merito alle quali si riporta una breve descrizione.

– Latino classico:

I dati sono tratti dal corpus *LASLA*[3], un'ampia raccolta di circa 1,7 milioni di parole provenienti da oltre 130 testi classici e tardo-latini lemmatizzati e contrassegnati morfologicamente. I dati di partenza sono stati resi liberamente disponibili presso il *repository* del laboratorio *LASLA* prima dell'inizio di questo progetto e sono accessibili attraverso la *LiLa Knowledge Base* all'indirizzo[4].

– Latino tardo:

I dati saranno presi dal *Confessiones Corpus* accessibile attraverso la *LiLa Knowledge Base* all'indirizzo[5].

– Latino medievale

I dati saranno presi dal *corpus* annotato sintatticamente *Index Thomisticus Treebank*[6]: il più grande *corpus* sintatticamente annotato disponibile per il latino, che comprende più di 400.000 token (cioè occorrenze di parole) dalle opere di Tommaso d'Aquino, compreso l'intero testo della *Summa contra Gentiles*.

L'annotazione viene effettuata manualmente da due annotatrici, che annotano i testi in modo indipendente tramite la risorsa digitale C.A.T. (*Content Annotation Tool*, precedentemente conosciuto come the *CELCT Annotation Tool*) creato e già usato per annotazione di coreferenze testuali [3]. L'accordo tra le annotatrici viene valutato a intervalli regolari per esaminare le scelte condivise e divergenti e, se necessario, le linee guida vengono perfezionate di conseguenza.

I tipi di coreferenza sono stati selezionati da quelli forniti dalla GUM[7] (*The Georgetown University Multilayer Corpus*; e sono elencati qui di seguito:

- ana - anaphoric, a pronoun referring back to something: [the woman] <-ana-- [she]. This is automatically generated from the 'coref' type when the anaphor is a pronoun.
- cata - cataphoric, a pronoun referring forward to something: [it]'s impossible [to know] ([it]--cata->[to know]). Automatically generated when the first member of a chain is a non-accessible pronoun.
- lexical coref - all types of coreference, including lexical mention: [Obama] .... <-coref-- [President Obama]
- split antecedent: [John] met [Mary] <-bridge-- [They] took a table together (in these cases the anaphor has multiple antecedents, but coreference only applies between the last mention and all previous mentions).

Altri tipi di coreferenza, come le descrizioni appositive e anaforiche, non vengono annotati, sia per limiti di tempo sia perché i tipi di coreferenza scelti per l'annotazione nel progetto sono quelli che possono essere considerati più "oggettivi", cioè quelli la cui annotazione non differisce pesantemente da una specifica teoria linguistica all'altra. Questo rappresenta un valore aggiunto dei risultati del *corpus* latino annotato con coreferenza, che si rivela utile per gli studiosi che lavorano in contesti teorici diversi, così come per i filologi interessati all'analisi testuale stilistica.

Infine, sarà utile sottolineare che in C.A.T. sono stati creati dei tag per annotare in modo funzionale alcuni elementi utili per la successiva fase di modellizzazione: lo strumento permette di distinguere tra *entity* e *head entity,* ovvero la *content entity* alla base della catena o relazione di coreferenza. È inoltre prevista anche l'indicazione della direzione verso cui si articola la relazione tra gli elementi coinvolti nonché di eventuali asimmetrie e discontinuità tra le entità. È infatti possibile marcare i casi di:

- *split antecedent*, quando un'entità si riferisce a due *head entities*; per esempio, in (1a), "Marco" e "Giovanni" sono due entità diverse che però fungono da unico antecedente del pronome "loro";
- "uno a molti", quando la *head entity* è composta da più elementi, come in (1b), dove "lui" si riferisce alla sequenza di elementi "Michele Mario Rossi" che rimanda a un'unica entità;
- *discontinuous markables*, quando la *head entity* è composta di più elementi non consecutivi nella sequenza testuale, come in (1c), in cui "lei" si riferisce a "Carla Lombardini" nonostante l'interposizione di "non Francesco".

[3] http://web.philo.ulg.ac.be/lasla/textes-latins-traites/
[4] https://lila-erc.eu/data/corpora/Lasla/id/corpus
[5] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Confessiones
[6] http://itreebank.marginalia.it
[7] https://wiki.gucorpling.org/gum/entities

(1)

. *Marco e Giovanni vanno a casa. Loro ordinano una pizza.*
. *Michele Mario Rossi sta cominciando a parlare. Lui è considerato un leader del settore.*
. *Carla, non Francesco, Lombardini ha vinto la competizione: lei è stata bravissima.*

All'interno dell'ultima sezione, forniremo un esempio di un testo latino in cui sono state individuate diverse tipologie di coreferenze.

4. ESEMPI DEL LAVORO DI ANNOTAZIONE

In questa sezione vengono forniti due esempi del lavoro di annotazione condotto tramite la risorsa digitale C.A.T., di cui la Fig. 1 mostra l'interfaccia, così come si presenta alle annotatrici: nella parte superiore è a disposizione il testo, diviso per frasi (si vedano sulla sinistra le sigle S0, S1, S2...), in cui è possibile marcare le *head entities* (in viola) e le *entities* (in verde) ad esse legate; nella parte inferiore lo strumento esplicita le relazioni stabilite tra le entità.

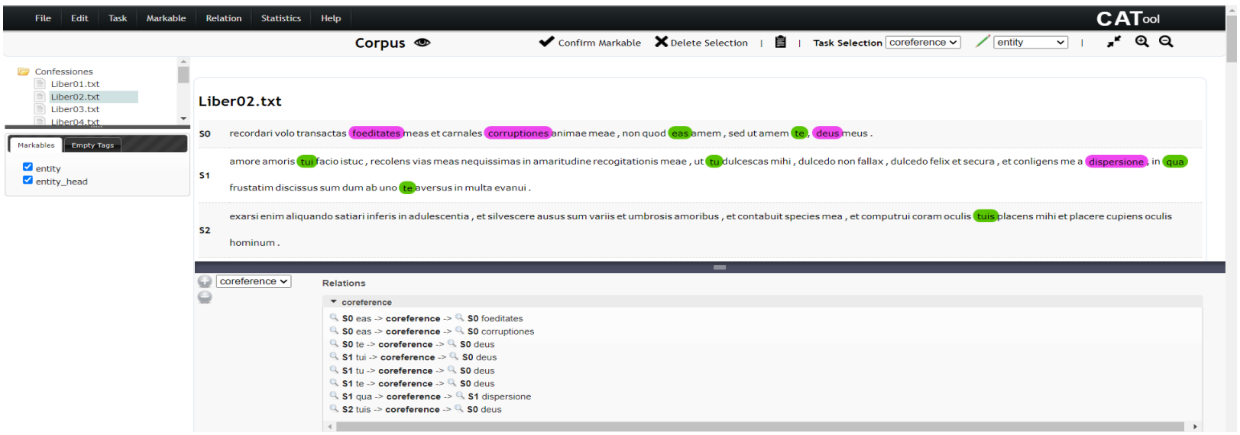


*Figura 1. Interfaccia C.A.T. (*Content Annotation Tool*)*

Il primo esempio testuale qui riportato è tratto dal secondo libro delle Confessioni di Agostino. Il testo di riferimento è collocato all'interno del *corpus* CIRCSE Latin Library.
Nel primo paragrafo del secondo libro sono state annotate tre relazioni di coreferenze, al fine di poterne descrivere la diversa struttura:

- 'foeditates' e 'corruptiones' sono *split anticedents* del pronome coreferente 'eas'.
- i pronomi e aggettivi 'tui', 'tu', 'te', tuis' costituiscono la catena di coreferenze di 'deus'.
- 'dispersione' è in una relazione di coreferenza binaria con 'qua'.

(2)
Recordari volo transactas foeditates meas et carnales corruptiones animae meae, non quod eas amem, sed ut amem te, deus meus. amore amoris tui facio istuc, recolens vias meas nequissimas in amaritudine recogitationis meae, ut tu dulcescas mihi, dulcedo non fallax, dulcedo felix et secura, et conligens me a dispersione, in qua frustatim discissus sum dum ab uno te aversus in multa evanui. exarsi enim aliquando satiari inferis in adulescentia, et silvescere ausus sum variis et umbrosis amoribus, et contabuit species mea, et computrui coram oculis tuis placens mihi et placere cupiens oculis hominum.

Il secondo esempio è invece tratto da una commedia, al fine di mostrare come il tipo di annotazione proposto può essere applicato a diversi generi testuali. Si tratta del *Curculio* di Plauto, disponibile presso il già citato *corpus* LASLA (cfr. *supra*). In (3) vediamo un passo dal I atto, in cui Fedromo e Palinuro discutono della schiava Planesio e del lenone di questa, Cappadoce. Possiamo notare anche in questo caso che diversi elementi sono messi in relazione tra loro:

- 'ei' (v. 43) risulta cataforico rispetto a 'lenoni' (v. 44);
- 'qui' (v. 44) risulta anaforico rispetto allo stesso 'lenoni';
- 'ancillula' (v. 43) rappresenta la *head entity* della catena coreferenziale che coinvolge 'eam', 'ea' (v. 46) e 'illa' (v. 47);
- 'me' (v. 46) è in una relazione cataforica con la *head entity* 'ego' (v. 47)

(3)

Phaed. At nunc veto.
sed ita uti occepi dicere: ei ancillula est.
Pal. Nempe huic lenoni qui hic habitat? **Phaed.** Recte tenes.
Pal. Minus formidabo, ne excidat. **Phaed.** Odiosus es.
Eam volt meretricem facere. Ea me deperit,
ego autem cum illa facere nolo mutuom.

RINGRAZIAMENTI

Questo contributo nasce in seno al progetto PRIN 2022 *Dati Testuali e Strumenti per la Risoluzione delle Coreferenze in Latino* (CUP J53D23013680008), in collaborazione fra il CIRCSE dell'Università Cattolica del Sacro Cuore di Milano e l'Università di Udine. Ringraziamo il Dottor Giovanni Moretti per la disponibilità e il supporto nell'uso dello strumento di annotazione.

BIBLIOGRAFIA

[1] Bach, Kent. *Thought and Reference*. Oxford: Clarendon Press, 1897.

[2] Bamman, David, e Patrick Burns. «Latin Bert: A Contextual Language Model for Classical Philology». https://doi.org/10.48550/arXiv.2009.10053 (2020).

[3] Bartalesi Lenzi, Valentina, Giovanni Moretti, e Rachele Sprugnoli. «CAT: the CELCT Annotation Tool». In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul 21-27 May 2012*, 333-338. European Language Resources Association (ELRA), 2012.

[4] Busa, Roberto. «Trent'anni d'informatica su testi: a che punto siamo? Quali spazi aperti alla ricerca?» In *L'Università e l'evoluzione delle Tecnologie Informatiche*, 1, 7:1-7.4, 1983.

[5] Calhoun, Sasha, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, e David Beaver. «The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue». *Language resources and evaluation* 44, fasc. 4 (2010): 387-419.

[6] Charniak, Eugene. *Towards a Model of Children's Story Comphrension*. Massachusetts: MIT, 1972.

[7] Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, e Weischedel Ralph. «The automatic content extraction (ACE) program – tasks, data, and evaluation». In *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon 26-28 May 2004*, 837-840. European Language Resources Association (ELRA), 2004.

[8] Grishman, Ralph, e Beth M. Sundheim. «Message understanding conference-6: A brief history». In *Proceedings of The16th International Conference on Computational Linguistics (COLING), Copenhagen 5-9 August 1996*, Vol. 1. Association for Computational Linguistics, 1996.

[9] Hirschman, Lyntte, Patricia Robinson, John Burger, e Marc Vilain. «Automating coreference: The role of annotated training data». In *Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing, Palo Alto 24-25 March 1997*, 1997.

[10] Kees van, Deemter, e Rodger Kibble. «On Coreferring: Coreference in MUC and Related Annotation Schemes». *Computational Linguistics* 26, fasc. 4 (2000): 629-637.

[11] Mitkov, Ruslan. *Anaphora Resolution.* London and New York: Routledge, 2022.

[12] Pradhan, Sameer, Lance Ramshaw, Marcus Mitchell, Martha Palmer, Ralph Weischedel, e Xue Nianwen. «CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes». In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Portland 11-19 June 2011*, 1-27, 2011.

Strumenti digitali per la trascrizione e la lemmatizzazione di testi in italiano antico

Emiliano Degl'Innocenti[1], Alessia Spadi[2], Federica Spinelli[3], Lucia Francalanci[4], Michela Perino[5], Irene Falini[6], Francesco Coradeschi[7], Francesco Pinna[8]

[1]CNR Istituto Opera del Vocabolario Italiano, Italia - emiliano.deglinnocenti@cnr.it
[2]CNR Istituto Opera del Vocabolario Italiano, Italia - alessia.spadi@cnr.it
[3]CNR Istituto Opera del Vocabolario Italiano, Italia - federica.spinelli@cnr.it
[4]CNR Istituto Opera del Vocabolario Italiano, Italia - lucia.francalanci@cnr.it
[5]CNR Istituto Opera del Vocabolario Italiano, Italia - michela.perino@cnr.it
[6]CNR Istituto Opera del Vocabolario Italiano, Italia - irene.falini@cnr.it
[7]CNR Istituto Opera del Vocabolario Italiano, Italia - francesco.coradeschi@cnr.it
[8] CNR Istituto Opera del Vocabolario Italiano, Italia - francesco.pinna@cnr.it

ABSTRACT

Il contributo si focalizza sullo sviluppo e sull'uso di metodologie per supportare e potenziare la ricerca nel contesto delle discipline umanistiche e del patrimonio culturale, con particolare riferimento all'ambito della filologia digitale. Partendo dal caso di studio del Fondo Datini dell'Archivio di Stato di Prato, l'obiettivo è lo sviluppo di nuovi strumenti digitali, nonché l'integrazione ed il potenziamento di strumenti esistenti, finalizzati allo studio del carteggio privato e commerciale del mercante pratese Francesco di Marco Datini. Lo scopo di questo progetto pilota è l'ampliamento, nel contesto del cluster H2IOSC [4], degli obiettivi raggiunti nell'ambito del progetto RESTORE (smaRt accESs TO digital heRitage and mEmory) in riferimento al trattamento di lettere edite che costituiscono il corpus lemmatizzato Archivio Datini realizzato dall'Istituto Opera del Vocabolario Italiano (OVI-CNR). L'implementazione di tali strumenti consentirà di facilitare la ricostruzione di una parte significativa della storia delle città d'Europa e dei porti del Mediterraneo del XIV secolo, evidenziandone sia le dinamiche della vita quotidiana, sia le specificità territoriali, sociopolitiche e commerciali.

PAROLE CHIAVE

HTR (Handwritten Text Recognition); lemmatizzazione; machine learning; filologia digitale; italiano antico.

1. INTRODUZIONE

Il progetto di ricerca descritto prende avvio dai risultati raggiunti dal progetto RESTORE[1] [3] (smaRt accESs TO digital heRitage and mEmory), incentrato sulla figura del mercante pratese Francesco di Marco Datini, della sua famiglia e del suo entourage di collaboratori. Partendo dalla dimensione locale è possibile ricostruire una parte rilevante della storia delle città dell'Europa e dei porti del Mediterraneo del XIV secolo, con le loro dinamiche sociolinguistiche e le loro peculiarità sociali ed economiche. Le due sezioni principali del Fondo Datini dell'Archivio di Stato di Prato sono infatti costituite dal carteggio privato tra Francesco e i suoi cari (tra cui la moglie Margherita Bandini e l'amico Lapo Mazzei) e dall'imponente carteggio commerciale che testimonia la fervida attività dei fondaci aziendali del mercante, situati in tutto il Mediterraneo.

Nello specifico, il punto di partenza di questo studio è il corpus (parzialmente) lemmatizzato Archivio Datini[2] – realizzato dall'Istituto Opera del Vocabolario Italiano del Consiglio Nazionale delle Ricerche (OVI-CNR) –, grazie al quale ci si propone di sviluppare una piattaforma di strumenti digitali integrati per supportare la ricerca nelle scienze umanistiche, focalizzandosi su alcune funzionalità chiave, quali la trascrizione e la lemmatizzazione automatica dei testi in italiano antico (con focus sulle scritture mercantesca e cancelleresca). L'obiettivo è fornire agli studiosi un ambiente avanzato per analizzare e interpretare i testi, che includa servizi integrati e interoperabili con strumenti già a disposizione della comunità dei ricercatori (quale ad esempio il TLIO).

Il contesto di riferimento è dato dagli studi condotti dal team riunito attorno al nodo italiano dell'Infrastruttura di Ricerca DARIAH[3] (Digital Research Infrastructure for the Arts and Humanities), con sede presso l'OVI[4], attualmente impegnato

[1] http://restore.ovi.cnr.it/
[2] Corpus lemmatizzato del carteggio Datini: http://aspweb.ovi.cnr.it/(S(acenhe55wjva14ulrxas2oyw))/CatForm01.aspx
[3] Nodo Italiano dell'Infrastruttura di Ricerca DARIAH: https://dariah.cnr.it/
[4] Istituto Opera del Vocabolario Italiano: http://www.ovi.cnr.it/

nel progetto H2IOSC[5] [4] (Humanities and cultural Heritage Italian Open Science Cloud), finanziato dal Piano Nazionale di Ripresa e Resilienza italiano (PNRR), che mira a creare un cluster partecipato dai nodi nazionali di 4 infrastrutture di ricerca ESFRI: DARIAH.it, CLARIN.it, OPERAS.it, E-RIHS.it. Nel corso degli ultimi anni, DARIAH.it ha indirizzato la sua attività verso la promozione dell'interoperabilità e l'accesso a risorse digitali legate al patrimonio culturale, sia tangibile (oggetti) che intangibile (elementi intellettuali e concettuali). Collaborando con comunità di ricerca consolidate, il gruppo si interfaccia con esperti del settore e promuove la Citizen Science al fine di plasmare lo sviluppo di un ecosistema digitale interoperabile per la ricerca nell'ambito delle Social Sciences and Humanities (SSH).

2. STATO DELL'ARTE

Nel corso del suo sviluppo e implementazione, il progetto RESTORE ha affrontato una serie di problematiche legate al trattamento dei dati, riassumibili in: 1) elevata frammentazione delle risorse digitali nei contesti di riferimento, che rischia di comprometterne il valore e ne limita la riutilizzabilità; 2) difficoltà di accesso e isolamento scientifico delle risorse di alta qualità prodotte da biblioteche, archivi e centri di ricerca; 3) eterogeneità dei formati e degli standard; 4) scarsa interoperabilità e carenza di strategie di sostenibilità a medio e lungo termine per le risorse digitali prodotte e gestite dagli attori coinvolti [8]. Inoltre, la crescente produzione di informazioni digitali accentua la sfida nell'organizzazione e nella gestione, mettendo in evidenza la necessità di criteri di selezione, strutturazione e pubblicazione dei dati per garantire qualità scientifica e interoperabilità. Il progetto, coordinato dall'OVI, ha coinvolto inizialmente istituzioni culturali del circuito GLAMs (Galleries, Libraries, Archives, Museums), a cui si sono uniti - nel corso del tempo - altri soggetti, attivi nel campo Social Sciences and Humanities (SSH) ed Heritage Science (HS) (ad esempio diagnostica e restauro del patrimonio culturale), con lo scopo di recuperare, integrare e rendere accessibili i dati, in linea con i principi FAIR[6] [16] (in breve: rintracciabilità, accesso, interoperabilità e riutilizzo). Per affrontare le sfide individuate è stato definito ed implementato un flusso di lavoro completo, capace di garantire l'integrazione e l'interoperabilità dei dati forniti da diversi soggetti, tra cui enti di ricerca, culturali e di conservazione nazionali e locali. Il flusso si articola nei seguenti passaggi: 1) acquisizione dei dati originali (nei formati in cui sono disponibili) forniti dai partner; 2) sviluppo di procedure per la normalizzazione e l'allineamento delle risorse codificate secondo gli standard dei domini di riferimento (ad es: TEI[7] per la codifica dei testi; EDM[8], MAG[9], MODS e METS[10] per la descrizione delle risorse bibliotecarie; EAD[11] e EAC[12] per la descrizione delle risorse archivistiche; ICCD[13] – in particolare la scheda OA – come sistema di catalogazione per le opere d'arte; altri standard afferenti a diverse discipline nel dominio delle scienze del patrimonio come EDF[14], HDF5[15] ecc.); 3) validazione dei dati normalizzati, attraverso la collaborazione con gli esperti di dominio; 4) mappatura e modellazione sulla base dell'ontologia scelta, CIDOC - Conceptual Reference Model[16]; 5) trasformazione dei dati in triple semantiche e caricamento nella base di dati semantica (Virtuoso Triplestore[17]); 6) esposizione di uno SPARQL endpoint per l'interrogazione della base di dati semantica e di un'interfaccia per la navigazione dei dati; 7) sviluppo di interfacce user-friendly, completamente integrate tra loro, per la visualizzazione dei dati, a cui si aggiunge l'uso di strumenti quali LodLive[18] (front-end per la visualizzazione grafica delle triple semantiche e la navigazione concettuale), EVT[19] (strumento open source per la progettazione e visualizzazione di edizioni digitali), Movio[20] (piattaforma open source multifunzionale per realizzare mostre virtuali). Inoltre, tutto il codice e la documentazione prodotti sono open source e archiviati in repository pubblicamente accessibili.

[5] https://www.h2iosc.cnr.it/
[6] FAIR: https://www.go-fair.org/fair-principles/
[7] Text Encoding Initiative - TEI: https://tei-c.org/
[8] Europeana Data Model - EDM: https://pro.europeana.eu/page/edm-documentation
[9] Administrative and Management Metadata - MAG: https://www.iccu.sbn.it/export/sites/iccu/documenti/manuale.html
[10] Metadata Object Description Schema - MODS e METS: https://www.loc.gov/standards/mods/presentations/mets-mods- morgan-ala07
[11] Encoded Archival Description - EAD: https://www.loc.gov/ead/
[12] Encoded Archival Context - EAC: https://eac.staatsbibliothek-berlin.de
[13] Istituto Centrale per il Catalogo e la Documentazione - ICCD: http://www.iccd.beniculturali.it/
[14] European Data Format - EDF: https://www.edfplus.info/
[15] Hierarchical Data Format - HDF5: https://www.hdfgroup.org/solutions/hdf5
[16] CIDOC - Conceptual Reference Model: http://www.cidoc-crm.org/
[17] Virtuoso Openlink Triplestore: https://virtuoso.openlinksw.com/
[18] LodLive: http://lodlive.it/
[19] Edition Visualization Technology – EVT: http://evt.labcd.unipi.it/
[20] Movio: https://www.gruppometa.it/it/movio

A partire dalla base di dati del progetto si vuole espandere il range di strumenti a disposizione di studiosi di varie discipline (ad es.: paleografi, filologi, lessicografi, linguisti, storici, filosofi ecc.) che si occupano di testi in italiano antico. Pertanto, DARIAH.it sta lavorando sia al potenziamento e alla FAIRificazione degli strumenti già realizzati nel contesto di RESTORE, sia allo sviluppo di strumenti di trascrizione e di lemmatizzazione (inclusa l'integrazione e la generalizzazione di strumenti esistenti) per varietà storiche di italiano. I servizi di trascrizione e lemmatizzazione sono in corso di addestramento su un dataset già trattato dal gruppo di lavoro e messo a disposizione dall'OVI e dall'Archivio di Stato di Prato per il progetto RESTORE, il già citato corpus testuale Archivio Datini, che raccoglie una selezione di lettere appartenenti al fondo omonimo, fisicamente depositato presso l'Archivio di Stato di Prato. Vista l'eterogeneità degli scriventi, il corpus comprende più varietà linguistiche e registra diverse forme grafiche e morfologiche di molti termini rilevanti per la ricostruzione del lessico quotidiano dell'epoca e del lessico tecnico legato alle attività economiche delle aziende datiniane. La lemmatizzazione approntata dai ricercatori dell'OVI include anche antroponimi, compresi eventuali soprannomi e posizioni specifiche (se l'indicazione si riferisce a una precisa personalità storica identificata), e toponimi, compresi nomi di città, paesi, distretti, località, strade, piazze, porte, chiese, monasteri, palazzi, ospedali, organizzazioni, istituzioni, ecc. Sono annotati inoltre termini relativi al campo religioso e agricolo, alle parti del corpo, alle scansioni temporali, ecc., distribuiti in 22 categorie concettuali (chiamate iperlemmi), tra cui: abbigliamento e arredamento, cibo, animali, arti e mestieri, calendario, legge ed economia politica, costruzione e architettura, medicina, monete, navigazione, parentela, cuoio e tessuti, e così via. In sintesi, il corpus è composto da: 2.511 testi; 45.259 forme; 977.034 occorrenze di cui 126.663 lemmatizzate; 6.510 lemmi e 22 iperlemmi (utilizzati per raggruppare diversi lemmi).

3. TRASCRITTORE

Nell'ambito del progetto è stato selezionato un dataset di circa 300 lettere appartenenti al corpus Archivio Datini, ciascuna associata alla rispettiva trascrizione e riproduzione digitale. L'associazione è stata resa possibile grazie al lavoro di integrazione e allineamento di dati, metadati e immagini effettuato durante la costruzione della base di dati semantica. Questa raccolta è stata scelta come caso di studio per addestrare uno strumento finalizzato al riconoscimento e alla trascrizione automatica del testo. Le lettere, redatte in scrittura mercantesca da diversi mittenti, costituiscono parte del carteggio commerciale e privato di Francesco Datini, così come precedentemente descritto. Il progetto di ricerca si propone di implementare un sistema di HTR (Handwritten Text Recognition), mirato alla trascrizione automatica della scrittura mercantesca. L'HTR utilizza modelli di apprendimento automatico, come reti neurali artificiali, per estrarre e interpretare caratteri scritti a mano in immagini, trasformandoli in testo digitale [1, 2, 5, 12]. Negli ultimi anni, l'HTR si è affermato come modello predominante nello sviluppo di strumenti, tra i quali, Loghi[21], eScriptorium[22] e Transkribus[23], dedicati alla trascrizione automatica di testi antichi. Nell'ambito di questo lavoro, il gruppo di ricerca DARIAH.it attivo presso l'OVI ha avviato una valutazione degli strumenti esistenti e del loro utilizzo da parte della comunità di ricerca di riferimento che ha portato all'individuazione del software eScriptorium - gratuito e open source, basato sull'OCR engine Kraken, anch'esso gratuito e open source - come uno dei modelli di riferimento per il progetto pilota in corso di sviluppo.

Il procedimento per la trascrizione automatica di testi comunemente comprende diverse fasi: preelaborazione dell'immagine, segmentazione, OCR/HTR, e post-elaborazione. La fase di segmentazione mira a individuare le linee di testo, preparandole per il successivo processo di trascrizione. Per questa fase - con l'obiettivo di addestrare un modello di segmentazione sperimentando l'utilizzo di eScriptorium, di cui è stata approntata una istanza locale - è stato selezionato un gruppo di lettere che presentano similitudini nel layout della pagina. Una delle sfide principali del progetto in corso di sviluppo risiede nella corretta gestione della variabilità dello stile di scrittura degli autori e nella tipologia corsiva della mercantesca; pertanto, il corpus di addestramento del carteggio Datini è stato selezionato anche tenendo conto del numero elevato di mani e mittenti.

Il progetto pilota fornirà un sistema ottimizzato per il riconoscimento del segno grafico e la successiva trascrizione assistita della scrittura, con particolare riferimento al carteggio considerato. Il sistema darà inoltre la possibilità di annotare la trascrizione in base a diverse tipologie di criteri (paleografico, filologico, storico, ecc.).

4. LEMMATIZZATORE

Questa parte del contributo è dedicata alla descrizione di alcuni esperimenti relativi all'annotazione linguistica automatica (in particolare *Part-of-Speech tagging* e lemmatizzazione) di varietà storiche di italiano. Proprio per la complessità dello

[21] https://github.com/rvankoert/loghi
[22] https://gitlab.com/scripta/escriptorium
[23] https://readcoop.eu/transkribus/

studio dell'italiano antico la scelta dei testi su cui operare è di fondamentale importanza quale requisito preliminare per procedere alla lemmatizzazione. I materiali scelti vengono organizzati in:

- Corpus di addestramento: uno o più testi, intesi come collezione di frasi compiute, lemmatizzati in modo esaustivo. Si considerano parte di questo corpus sia i testi usati per l'addestramento che quelli usati per la valutazione.
- Corpus di lingua (o lessico): insieme di testi che comprenda il corpus di addestramento, ma con numero totale di occorrenze molto maggiore. Il corpus di lingua non deve essere lemmatizzato esaustivamente ma è preferibile (anche se non indispensabile) che sia lemmatizzato esaustivamente per forme. In particolare, per l'esperimento che si sta descrivendo si prende come riferimento il TLIO (Tesoro della Lingua Italiana delle Origini)[24] dove tutte le forme sono lemmatizzate almeno una volta, anche se non in tutte le occorrenze.

Alla costruzione dei corpora si è affiancata un'analisi esplorativa sullo stato dell'arte delle risorse e degli strumenti dedicati al trattamento automatico dell'italiano antico [6, 7, 13, 14, 15], da cui sono emerse due possibili tipologie di approcci: da una parte, l'uso di strumenti specifici per l'italiano antico, dall'altra lo sviluppo o il riadattamento di strumenti di annotazione addestrati sull'italiano contemporaneo.

La lemmatizzazione procede per "frasi", dove per "frase" si intende la parte di testo tra l'inizio o un segno di interpunzione forte e la fine o il segno di interpunzione forte successivo, ovvero quello che si definisce un periodo. Una volta individuate le frasi la procedura si può riassumere nei seguenti punti: 1) per ogni frase si considerano tutte le occorrenze; 2) tramite il lessico si associano ad ogni occorrenza tutti i lemmi a cui corrispondono; 3) si costruiscono le catene di lemmi della frase. Una catena di lemmi consiste in una sequenza di due o più lemmi (lo standard si attesta su catene di cinque lemmi ma non è dogmatico), intese come terne costituite da forma standard, categoria grammaticale e disambiguatore. Per ogni occorrenza viene costruito un insieme di catene di lemmi in tutte le combinazioni disponibili nel lessico già indentificate al punto 2; 4) ad ogni catena di lemmi si associa una probabilità intrinseca calcolata sulla base della probabilità forma/lemma e della probabilità lemma/struttura; 5) una volta note le probabilità intrinseche di tutte le catene di lemmi della frase si stima per l'intera frase la concatenazione di lemmi di massima probabilità. Per determinare la sequenza di lemmi di massima probabilità potrebbe sembrare necessario calcolare la probabilità di tutte le combinazioni di catene di lemmi possibili, con relativo costo computazionale piuttosto alto. In realtà si vede che partendo da un lato della frase (l'inizio o la fine è equivalente) ed aggiungendo di volta in volta un "anello" alla catena, stimando le probabilità tramite l'algoritmo di Viterbi[25], si riesce a mantenere la complessità algoritmica della procedura sotto controllo.

L'obiettivo dello studio è la realizzazione di un servizio per la lemmatizzazione semiautomatica di testi scritti in italiano antico basato sul machine learning. La realizzazione di un tale servizio è strettamente collegata alla costruzione di uno o più corpora annotati rappresentativi delle varietà storiche, da poter usare in fase di addestramento e di valutazione.

Uno di questi dataset è costituito dal sotto corpus creato a partire dalle lettere appartenenti all'Archivio Datini, precedentemente descritto. Gli esperimenti condotti su questi testi hanno evidenziato sia le problematiche connesse alle peculiarità specifiche di testi antichi, come l'alta variabilità a tutti i livelli di analisi linguistica (grafico, morfologico, sintattico e lessicale), sia i problemi derivanti dall'adattamento di strumenti preesistenti a varietà linguistiche differenti. Una parte dei test è stata effettuata sul sistema di lemmatizzazione semiautomatica realizzato nel 2010 da Domenico Iorio-Fili [9, 10] e inserito all'interno della versione 4.0 di GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini), il software che gestisce la versione elettronica dei corpora testuali dell'OVI. Si tratta di uno strumento propriamente costruito per il trattamento dell'italiano antico e in particolare per il caso specifico del Corpus OVI, caratterizzato da dimensioni relativamente ridotte, ma da un'elevata complessità e variabilità del materiale linguistico. Altre prove sono state invece condotte con strumenti di NLP (Natural Language Processing) addestrati sull'italiano contemporaneo e sviluppati in ambiente Python.

5. CONCLUSIONI

Visti i riscontri positivi da parte della comunità scientifica circa le possibilità offerte da RESTORE per la ricostruzione del lessico e della fitta rete di persone e luoghi gravitanti attorno alla figura di Francesco di Marco Datini nel Mediterraneo del XIV sec., l'Archivio Datini si configura come l'oggetto di studio ideale per testare l'efficacia, anche in termini di riutilizzabilità, di due strumenti, trascrittore e lemmatizzatore, indubitabilmente utili alla ricerca scientifica in molteplici discipline che hanno come punto di partenza l'interpretazione di testi in italiano antico. Lo sviluppo di queste tecnologie da parte di DARIAH.it si inserisce nella progettazione di uno strumento pilota dedicato alla filologia digitale (Digital

[24] http://tlio.ovi.cnr.it/TLIO/

[25] L'algoritmo Viterbi è un algoritmo ideato da Andrew Viterbi e generalmente utilizzato per trovare la migliore sequenza di stati (detta Viterbi path) in una sequenza di eventi osservati in un processo markoviano. Da Wikipedia: https://it.wikipedia.org/wiki/Algoritmo_di_Viterbi

Philology Hub) (vd. Fig. 1) previsto in seno a H2IOSC, per l'ideazione della quale si stanno seguendo le linee guida esposte in Leonardi (2021) [11] a lungo discusse con il gruppo DARIAH.it attivo presso l'OVI. La progettazione di questo pilot per la ricerca filologica rientra inoltre tra gli obiettivi della collaborazione fra lo Spoke 3 del progetto CHANGES[26] («Digital library, archives and philology») ed H2IOSC, con particolare riferimento all'attività di sviluppo del Digital Philology Hub, coordinata dall'OVI per DARIAH.it; nel medesimo contesto di collaborazione si colloca l'istituzione del corso di dottorato FROID[27] («Filologia Romanza e Italiana Digitale») presso la Scuola Normale Superiore[28], che vede la compartecipazione di DARIAH.it e dell'OVI attraverso il finanziamento di una borsa di dottorato legata allo sviluppo dell'Hub: primi esempi della fruttuosa sinergia tra i progetti PNRR IR H2IOSC e PE CHANGES.

Figura 1. Il Digital Philology Hub: le fasi del lavoro filologico descritte in L. Leonardi, "Filologia digitale del Medioevo italiano", pubblicato in Italianistica digitale = «Griseldaonline», 20, 2 (2021), pp. 77-89.

6. RINGRAZIAMENTI

Progetto H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 4 "Istruzione e Ricerca" Componente 2 "Dalla ricerca all'impresa" Linea di Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione" Azione 3.1.1 "Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti" - Codice progetto IR0000029 - CUP B63C22000730005. Soggetto attuatore CNR.

[26] CHANGES: https://sites.google.com/uniroma1.it/changes/home
[27] FROID: https://www.sns.it/it/disciplinacorso-di-laurea/corso-phd/filologia-romanza-e-italiana-digitale-froid
[28] SNS: https://www.sns.it/it

BIBLIOGRAFIA

[1] Cascianelli, Silvia, Marcella Cornia, Lorenzo Baraldi, Maria Ludovica Piazzi, Rosiana Schiuma, and Rita Cucchiara. "Learning to Read L'Infinito: Handwritten Text Recognition with Synthetic Training Data." In *Computer Analysis of Images and Patterns. CAIP 2021*, edited by Nicolas Tsapatsoulis, Andreas Panayides, Theo Theocharides, Andreas Lanitis, Constantinos Pattichis, and Mario Vento, 13053:340–350. Lecture Notes in Computer Science. Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-89131- 2_31

[2] Clérice, Thibault, Malamatenia Vlachou-Efstathiou, and Alix Chagué. "CREMMA Medii Aevi: Literary Manuscript Text Recognition in Latin." *Journal of Open Humanities Data* 9 (2023): 1–19. https://doi.org/10.5334/johd.97

[3] Coradeschi, Francesco, Leonardo Canova, Emiliano Degl'Innocent, Carmen Di Meo, Maurizio Sanedi, Alessia Spadi, and Federica Spinelli. "The RESTORE Project: A Final Review." In *Proceedings of the 19th Conference on Information and Research Science Connecting to Digital and Library Science*, edited by Alessia Bardi, Alex Falcon, Stefano Ferilli, Stefano Marchesin, and Domenico Redavid, 167–179. Bari, 2023.

[4] Degl'Innocenti, Emiliano, Monica Monachini, Alberto Bucciero, Enrico Pasini, Bruno Fasini, and Francesca Frontini. "H2IOSC: Humanities and Heritage Open Science Cloud." In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, edited by Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 63-64, 2023. https://doi.org/10.6092/unibo/amsacta/7721

[5] Dhiaf, Marwa, Ahmed Cheikh Rouhou, Yousri Kessentini, and Sinda Ben Salem. "MSdocTr-Lite: A Lite Transformer for Full Page Multi-Script Handwriting Recognition." *Pattern Recognition Letters* 169 (2023): 28–34. https://doi.org/10.1016/j.patrec.2023.03.020

[6] Favaro, Manuel, Marco Biffi, and Simonetta Montemagni. "Risorse linguistiche di varietà storiche di italiano: il progetto TrAVaSI." In *Proceedings of the Seventh Italian Conference on Computational Linguistics. CLiC-It 2020 (Bologna, Italy, March 1-3, 2021)*, edited by Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, 178–86. Torino: Accademia University Press, 2020.

[7] Favaro, Manuel, Marco Biffi, and Simonetta Montemagni. "Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione." In *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data, JADT 2022*, edited by Michelangelo Misuraca, Germana Scepi, and Maria Spano, 1:392–399. Napoli: Vadistat press, 2022.

[8] Hilbert, Martin. "How Much Information Is There in the 'Information Society'?" *Significance* 9, no. 4 (2012): 8–12.

[9] Iorio-Fili, Domenico. "Il Lemmatizzatore Semiautomatico Di GATTO4." In Dizionari e Ricerca Filologica, Atti Della Giornata Di Studi in Memoria Di Valentina Pollidori. Bollettino Dell'Opera Del Vocabolario Italiano, Supplemento III:41–56, 2010.

[10] Iorio-Fili, Domenico. "Un Nuovo Strumento Di Lemmatizzazione Automatica per Corpora Testuali Di Ridotte Dimensioni. Applicazione All'italiano Antico." *Bollettino Dell'Opera Del Vocabolario Italiano* XV (2010): 367–391.

[11] Leonardi, Lino. "Filologia Digitale Del Medioevo Italiano." *Italianistica Digitale, Griseldaonline* XX, no. 2 (2021): 77–89.

[12] Lombardi, Francesco, and Simone Marinai. "Deep Learning for Historical Document Analysis and Recognition - A Survey." *Journal of Imaging* 6, no. 10 (2020): 110. https://doi.org/10.3390/jimaging6100110

[13] Montemagni, Simonetta. "Trattamento automatico del linguaggio e Digital Humanities: metodi e strumenti, sfide." In *Digital Humanities. Metodi, strumenti, saperi*, edited by Fabio Ciotti, 160–177. Roma: Carocci, 2023.

[14] Pennacchiotti, Marco, and Fabio M. Zanzotto. "Natural Language Processing Across Time: An Empirical Investigation on Italian." In *Proceedings of GoTAL - 6th International Conference on Natural Language Processing, LNAI 5221*, edited by Bengt Nordström and Aarne Ranta, 5221:371–382. Lecture Notes in Computer Science. Gothenburg: Springer, 2008.

[15] Piotrowski, Michael. "Natural Language Processing for Historical Texts." *Synthesis Lectures on Human Language Technologies* 5, no. 2 (2012): 1–157. https://doi.org/10.2200/S00436ED1V01Y201207HLT017

[16] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

Testi allografici: contatti tra lingue e scritture del mediterraneo

Antonio Pagliara[1], Federico Boschetti[2], Daniele Baglioni[3]
[1] Università Ca' Foscari di Venezia, Italia - antonio.pagliara@unive.it
[2] Università Ca' Foscari di Venezia, Italia - federico.boschetti@unive.it
[3] Università Ca' Foscari di Venezia, Italia - daniele.baglioni@unive.it

ABSTRACT
Questo contributo discute la creazione dell'edizione scientifica digitale e l'analisi linguistica di testi allografici italoromanzi, documenti unici scritti in volgari italiani ma con alfabeti non latini dell'area mediterranea quali il greco, l'ebraico, l'arabo e il siriaco. Nonostante sia un fenomeno noto, ha ricevuto scarsa attenzione nella ricerca storico-linguistica e filologica in Italia, principalmente a causa delle difficoltà interpretative. Il progetto MIA (Manuscripta Italica Allographica) e altre iniziative simili mirano a catalogare digitalmente questi testi, applicando i metodi della filologia digitale per trascriverli e interpretarli. L'analisi si concentra sulle peculiarità grafiche e fonologiche, evidenziando l'influenza dei sistemi grafici originali sulla lingua italiana scritta e offrendo nuove prospettive sulla competenza linguistica degli autori alloglotti. Questa ricerca apporta un contributo significativo alla comprensione della diversità grafica e linguistica dei testi allografici, proponendo metodologie innovative per il loro studio attraverso l'apporto delle Digital Humanities.

PAROLE CHIAVE
Allografia; filologia digitale; contatto linguistico; interferenza linguistica; chirurgia medievale.

1. INTRODUZIONE
Il presente contributo riguarda l'edizione digitale e l'analisi linguistica di testi allografici italoromanzi, cioè di documenti nei quali i volgari italiani sono stati resi con sistemi grafici differenti da quello latino, come l'alfabeto greco ed ebraico o, in esemplari isolati, arabo e siriaco. Il fenomeno è noto agli studiosi da tempo, almeno dalla seconda metà dell'Ottocento, e tuttavia è rimasto ai margini delle ricerche storico-linguistiche e filologiche in Italia, probabilmente per le difficoltà poste dall'edizione e dall'interpretazione di scritture "altre". Esso merita tuttavia attenzione, anzitutto per la varietà delle scritture impiegate, che non ha eguali nel resto della Romània: mentre infatti il ricorso alle scritture semitiche è, com'è noto, ben attestato soprattutto in area iberoromanza, l'uso dei caratteri greci è invece esclusivo dell'Italia. Anche la quantità non deve essere sottovalutata: come ha notato recentemente Rubin [11], la tradizione di testi italiani in scrittura ebraica, che si stima intorno ai 200 documenti e comprende anche importanti testimonianze letterarie, è di gran lunga la più abbondante in ambito romanzo, con la sola eccezione del giudeo-spagnolo; il corpus in alfabeto greco è più contenuto (Basile [1] elenca una cinquantina di testi), ma presenta un maggior numero di tradizioni scrittorie, distribuite fra il Meridione estremo, da cui proviene la gran parte dei documenti, la Sardegna e il Veneto [10]. Limitata a due soli esempi è invece la documentazione rispettivamente in scrittura araba [9] e siriaca [6], che pone questioni ancora irrisolte da una parte sulle funzioni dell'allografia e sui destinatari, dall'altra sulle modalità della "transcritturazione" (sul concetto, cfr. [7: 20-21]), specie se confrontata con analoghi testi in altre lingue romanze (per l'arabo) e non romanze (per il siriaco).
Come si evince facilmente dalle date di pubblicazione della bibliografia indicata, negli ultimi decenni si è assistito a una notevole ripresa degli studi allografici [8], che ha portato alla scoperta di testi nuovi (per alcuni esempi in caratteri greci cfr. [3, 4, 5]) e anche all'applicazione alla documentazione nota degli strumenti della filologia digitale. In quest'ultimo ambito si collocano vari progetti recentemente finanziati. Anzitutto, va menzionato MIA - Manuscripta Italica Allographica (PRIN 2022), che vede coinvolte le Università di Pisa, Venezia Ca' Foscari, Napoli "Federico II" e Messina e il CNR - Istituto di Linguistica Computazionale "Antonio Zampolli" di Pisa. Obiettivo del progetto, è di fornire il primo catalogo digitale di tutti i testi allografici manoscritti noti in ambito italoromanzo, nonché di sperimentare l'edizione digitale di alcuni brevi documenti in caratteri greci ed ebraici. A MIA si affianca l'attività dell'unità veneziana all'interno dello *spoke* 3 (WP 4) del progetto PNRR CHANGES (*leader* Università "Federico II" di Napoli, *co-leader* Università di Bergamo), che sarà dedicata all'edizione del testo di gran lunga più esteso in scrittura greca, il quattrocentesco volgarizzamento della *Chirurgia* di Guglielmo da Saliceto [2] (per l'edizione del proemio dell'opera cfr. [10]). Dei cinque libri che compongono il trattato, il quinto sarà reso disponibile in edizione digitale, e consentirà, attraverso un sistema automatico di traslitterazione e la sperimentazione di un sistema computer-assisted per la trascrizione interpretativa, uno

studio *computer-based* delle modalità di transcritturazione del volgare italiano in alfabeto greco, con auspicabili ricadute metodologiche su tutte le ricerche inerenti alle allografie. Uno specime del lavoro, che dà bene l'idea delle potenzialità di questo studio, condotto con il sostegno del Venice Centre for Digital and Public Humanities (VeDPH) del Dipartimento di Studi Umanistici di Ca' Foscari, è offerto in § 4.

2. METODI ECDOTICI

Dando priorità ai materiali già disponibili online con licenza aperta, si procede al caricamento delle immagini digitali su *eScriptorium*. Si è privilegiata questa piattaforma web di Handwritten Text Recognition (HTR) per la sua adesione ai principi della scienza aperta in tutte le fasi del processo di digitalizzazione, infatti, non solo il codice sorgente di *eScriptorium* e di *kraken* (il motore HTR su cui *eScriptorium* si basa) sono liberi, ma anche i modelli HTR, i documenti usati per produrli e i relativi metadati sono resi disponibili con licenze aperte, grazie al progetto HTR United[1].

Tramite *eScriptorium* viene condotta la layout analysis delle pagine dei manoscritti, in modo da poter eseguire la mappatura del testo sull'immagine per il confronto diretto con il fac-simile digitale, linea per linea. I vari blocchi di testo vengono inoltre classificati manualmente secondo le linee guida del progetto SegmOnto[2], al fine di identificare sulla pagina i differenti flussi testuali (ad es. colonne di testo) e paratestuali (ad es. glosse o marginalia). Solo per la *Chirurgia*, composta da 158 carte, si è pianificato di trascrivere manualmente per intero un solo libro (il quinto, di 34 carte), con il proposito di creare un modello HTR da applicare alle carte restanti, sulle quali calcolare, a campione, le prestazioni del sistema di riconoscimento automatico e procedere in un secondo momento alla correzione manuale. Per tutte le altre testimonianze, data l'estensione notevolmente più ridotta, si è deciso invece di procedere alla sola trascrizione manuale, sfruttando *eScriptorium* soltanto per la layout analysis.

La trascrizione dei testi viene eseguita usando un Domain-Specific Language creato per il progetto (MIADSL) per la codifica dei fenomeni testuali quali, a titolo di esempio, abbreviazioni, lettere sovrascritte, espunzioni del copista o presenza di marginalia. La context-free grammar che definisce il lessico e la sintassi di MIADSL è stata studiata dai filologi digitali insieme ai collaboratori del progetto con un approccio più tradizionale, al fine di garantire una maggiore leggibilità dei documenti codificati. Tramite parsing del DSL con ANTLR[3] che produce file XML con schema proprietario, e fogli di stile XSLT, i documenti vengono convertiti in XML/TEI.

Considerando la forte interdisciplinarità richiesta dallo studio di sistemi grafici molto diversi fra loro (greco, ebraico, arabo e siriaco), la traslitterazione è utile sia agli specialisti del progetto, sia al più vasto pubblico dei filologi romanzi. Questi documenti, infatti, oltre ad avere un valore intrinseco per la loro peculiare veste grafica, sono di grande interesse anche per la linguistica storica, la sociolinguistica, nonché la storia della lingua e la storia della letteratura. Si è scelto di adottare un sistema di traslitterazione biunivoca per far corrispondere a ciascun grafema (digramma o trigramma) della scrittura originale un solo grafema (digramma o trigramma) in caratteri latini. La traslitterazione, quindi, viene eseguita in modo completamente automatico.

Oltre alla semplice traslitterazione, è necessaria una trascrizione interpretativa in alfabeto latino per comprendere la rifunzionalizzazione degli originali sistemi di scrittura adottati nelle diverse opere per la resa dei volgari italiani. Questa operazione procede quindi alla decodifica delle strategie di transcritturazione [7] messe in atto dagli autori o dai copisti. Si prenda ad esempio ντὲφέντα, traslitterato automaticamente in *ntèféntα* e interpretato dagli studiosi del progetto come *defenda*. Come si può vedere, in questo caso la prima occorrenza del digramma *nt-* ha funzione di *-d-* e la seconda occorrenza ha funzione di *-nd-*, come ci si aspetterebbe in greco moderno.

L'edizione digitale di queste opere è data quindi dall'insieme della trascrizione semidiplomatica, della traslitterazione e della trascrizione interpretativa in alfabeto latino, messe in relazione fra di loro a granularità di parola (o, più precisamente, di token).

3. METODI DI ANALISI LINGUISTICA

Lo studio linguistico di questi testi ha lo scopo di analizzare le interferenze fra sistemi grafemici e sistemi fonologici delle diverse lingue prese in considerazione. Si vuole indagare quale sia il diverso grado di competenza degli scriventi alloglotti nei confronti del volgare italiano: una elevata padronanza della lingua parlata non è necessariamente correlata a una piena padronanza della lingua scritta e viceversa. Da alcune spie grafiche e linguistiche è possibile ipotizzare che il sistema della

[1] https://htr-united.github.io
[2] https://segmonto.github.io
[3] https://www.antlr.org

lingua madre condizioni la resa dei volgari italiani, dimostrando allo stesso tempo una influenza non solo a livello grafemico ma anche morfosintattico.
Per questo i metodi adottati sono sia di tipo quantitativo che qualitativo. Tramite l'analisi statistica è possibile, ad esempio, verificare la produttività dei diversi allografi che concorrono a esprimere uno stesso fonema del volgare italiano (si pensi ai molti modi di rappresentare la *i* a cui ricorre il sistema di scrittura del greco bizantino e moderno). Ma a volte è solo mediante l'analisi qualitativa che è possibile ascrivere fenomeni che si manifestano molto raramente o addirittura una sola volta a una specifica varietà linguistica (si pensi ad esempio alla presenza di venezianismi lessicali in un'opera come la *Chirurgia* allografica).
Per l'analisi quantitativa, dalla trascrizione semidiplomatica di ciascuna parola viene generata automaticamente una regular expression che rende conto delle sue possibili realizzazioni fonologiche, mentre dalla trascrizione interpretativa in alfabeto latino viene generata la sequenza di fonemi (o arcifonemi) ipotizzata per la sequenza di grafemi che rappresenta quella specifica forma lessicale. Un esempio dovrebbe chiarire meglio il procedimento. Si prenda la trascrizione semidiplomatica κείντω (*keíntō*) e la relativa trascrizione interpretativa «*chinto*», non attestata altrove ma prossima alla grafia di *chuinto*, che si trova in TLIO[4] con il significato di *quinto*. Dalla trascrizione semidiplomatica si ricava quindi la regular expression r"ke?in?[dt]O", che esprime la possibilità per il dittongo -εί- di essere realizzato come *-ei-* oppure *-i-* e la possibilità per il nesso -ντ- di essere realizzato in ben quattro modi diversi: *-nd-*, *-nt-*, *-d-* oppure *-t-*. Dalla trascrizione interpretativa si ottiene invece la stringa "kintO" (con O maiuscola da intendere come arcifonema di o aperta oppure chiusa). L'allineamento dei vari segmenti della regular expression e della stringa (r"k"→"k", r"e?i"→"i", r"n?[dt]"→"nt", r"O"→"O") permette di studiare le differenti realizzazioni degli allografi.
Per l'analisi qualitativa, si sta mettendo a punto invece un DSL che permetta di annotare in modo sintetico rilevanti fenomeni morfosintattici e lessicali, al fine di indicizzarli, raggrupparli e mostrarli in contesto. Il venezianismo lessicale μπάντα (*mpánta*), in trascrizione interpretativa «*banda*», viene annotato sinteticamente nel seguente modo: μπάντα : banda (vec) ≈ lato (it).

Anche le annotazioni linguistiche realizzate tramite DSL vengono convertite in TEI (secondo le linee guida relative alla stand-off annotation e all'analisi linguistica). Tutte le risorse digitali prodotte saranno poi depositate per la long-term preservation sul repository di ILC4CLARIN[5] in ottemperanza ai principi FAIR e sfruttando l'infrastruttura di H2IOSC[6].

4. RISULTATI PARZIALI

Vale la pena vedere ora concretamente quali sono i primi risultati ottenuti, a partire da un passo della *Chirurgia*.
[129r[l].9-21] Trascrizione: σοῦπρα | ἀ κέστω πανύκουλο ἐ ὀρδυνάτω · ||10 οὔνω ἄλτρω πανύκουλο πίου ντού|ρου · καὶ κέστω ἐ ἀτζῶ καὶ ντὲφέντα · | λὸ τζερυέλω · ἐ λὸ πριμὲ πανύκου|λο ντὲ λε ντουρίτζιε · ντελ όσω · ἐ ἀ ||15 κέλω πανυκούλω · ἀνκούρα ἐ τὲ|σούτω · ντε βέναι · ἐν ντὲ ἀρταρί|αι : φέρμαι · αἰ εἰν ἐζέμπιο ντε μ|πωνυτάται · ντὲ σῶα · κὸμποζητι|ῶναι παρτικουλάρε · […]||20
Traslitterazione: Soýpra a késtō panýkoylo e ordynátō · oýnō áltrō panýkoylo píoy ntoýroy · kaì késtō e atzō kaì ntèfénta · lò tzeryélō · e lò primè panýkoylo ntè le ntoyrítzie · ntel ósō · e a kélō panykoýlō · ankoýra e tèsoýtō · nte vénai · en ntè artaríai : férmai · ai ein ezémpio nte mpōnytátai · ntè sōa · kompozītiōnai partikoyláre.
Trascrizione interpretativa: Supra a chesto paniculo è ordinato uno altro paniculo più duru che chesto e açò che defenda lo çervelo e lo primè paniculo de le duriçie del'oso. E a chelo paniculo ancura è tesuto de vene e de artarie ferme è in esempio de bonitate de soa compositione particulare.
Dal breve passo tratto dal IV libro della *Chirurgia* qui riportato, unitamente ad una sua traslitterazione e una prova di trascrizione interpretativa, si evince come il copista ricorra ad un esasperato poligrafismo e a un utilizzo ipertrofico dell'apparato paragrafematico greco. Dall'esame grafico-linguistico condotto sul IV libro dell'opera si è potuto dimostrare, con un certo margine di sicurezza, l'origine greca del copista alla luce della distribuzione dei grafemi vocalici (come i digrammi <οι> e <αι>), condizionata morfologicamente e lessicalmente dalla lingua greca, nonché dalla rappresentazione di /b/ e /d/ tramite i digrammi <μπ> e <ντ> di origine bizantina. Inoltre, sono riscontrabili diversi tratti di interferenza del greco che riflettono una sovrapposizione del sistema fonologico della lingua d'origine nella transcritturazione (si noti, nel paragrafo riportato, l'innalzamento di /o/ in /u/ in sede tonica e atona). L'antigrafo doveva essere chiaramente in caratteri latini come dimostra la presenza di calchi grafici di alcuni nessi latini, mentre la lingua dell'opera appare di più difficile definizione [1]. Se, infatti, è innegabile una patina veneta, e più specificatamente venezianeggiante, sono molto più

[4] http://tlio.ovi.cnr.it
[5] https://ilc4clarin.ilc.cnr.it
[6] https://www.h2iosc.cnr.it

frequenti i fenomeni linguistici di matrice tosco-fiorentina. Un'analisi completa dell'opera condotta con il sussidio di un quantificatore digitale delineerà con maggior precisione i processi alla base dei meccanismi di transcritturazione, da una parte indagandone le peculiarità (come, a titolo di esempio, la resa della fricativa /v/ tramite il grafema /υ/) e la loro incidenza nell'economia globale dell'opera, dall'altra interpretando e assegnando un valore fonologico a quei nessi grafici (come <τζ>) che possano sancire l'appartenenza del testo ad una determinata area linguistica.

5. CONCLUSIONE

L'applicazione della filologia digitale a testi in scrittura greca (e anche ebraica e araba) non è certo nuova, e ha già prodotto importanti risultati (si pensi, fra gli altri, al *Cologne Papyrus Portal* dell'Università di Colonia e al progetto *Greek into Arabic. Philosophical Concepts and Linguistic Bridges* del CNR-ILC di Pisa). Di recente, l'applicazione ha cominciato a essere estesa anche a tesi allografici in lingue romanze: è il caso del progetto *Bible Glossaries as Hidden Cultural Carriers* dell'Università di Heidelberg, dedicato ai glossari biblici in giudeo-francese. Tuttavia la documentazione allografica italoromanza, malgrado il suo interesse, non è stata ancora interessata da questo tipo di ricerche. I progetti elencati in § 1, la cui metodologia è stata illustrata nei paragrafi precedenti sulla base della *Chirurgia* volgare in caratteri greci, costituiscono pertanto un importante elemento di novità, su due diversi fronti. Da un lato, contribuiscono alla conoscenza, alla valorizzazione e allo studio scientifico di documenti in volgare che, per le proprie caratteristiche grafiche, mancano dalle principali banche dati online (in primis l'OVI). Dall'altro, in virtù della scelta di non concentrarsi su una scrittura sola e di confrontare piuttosto le diverse testimonianze allografiche, focalizzando l'attenzione sui processi di transcritturazione ed elaborando sistemi automatici di traslitterazione dei documenti, ambiscono a fornire una metodologia unica per lo studio di questi testi, al cui sviluppo le *digital humanities* concorrono in maniera determinante.

6. RICONOSCIMENTI

I risultati discussi sono stati elaborati nell'ambito del Progetto PRIN 2022 MIA "Manuscripta Italica Allographica. Italo-Romance Texts Written in non-Latin Characters from the Middle Ages to Modern Times" finanziato dall'Unione Europea – Next Generation EU nell'ambito del progetto PNRR M4C2 - Investimento 1.1 "Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) PROT. N. 2022ZAH9HC - CUP B53D23009850006, e del progetto "CHANGES - Cultural Heritage Active Innovation for Sustainable Society" finanziato dall'Unione Europea – Next Generation EU e dal Ministero dell'Università e della Ricerca su fondi PNRR (PE00000020 - CUP: H53C22000850006).
Le risorse e i servizi infrastrutturali sono messi a disposizione dal Progetto H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione europea NextGenerationEU – PNRR M4C2 - Codice progetto IR0000029 - CUP B63C22000730005.

BIBLIOGRAFIA

[1] Baglioni, Daniele. 'Altre scritture'. In *Storia dell'italiano scritto*, a cura di Giuseppe Antonelli, Matteo Motolese, and Lorenzo Tomasin, 6. Pratiche di scrittura:84–110. Roma: Carocci, 2021.

[2] Baglioni, Daniele. 'Italoromanzo in caratteri arabi in un diploma magrebino del Trecento'. In *Contatti di lingue – Contatti di scritture. Multilinguismo e multigrafismo dal Vicino Oriente Antico alla Cina contemporanea*, edited by Daniele Baglioni and Olga Tribolato, 177–197. Venezia: Edizioni Ca' Foscari, 2015.

[3] Baglioni, Daniele, and Olga Tribulato. Contatti di lingue – Contatti di scritture. Multilinguismo e multigrafismo dal Vicino Oriente Antico alla Cina Contemporanea. Edizioni Ca' Foscari: Venezia, 2015.

[4] Basile, Angela. 'Repertorio dei testi romanzi in caratteri greci dell'Italia Meridionale e della Sicilia (secc. XII-XVI)'. *Medioevo Letterario d'Italia* 9 (2012): 49–88.

[5] Coco, Alessandra, and Francesca Di Stefano. 'La «Chirurgia» di Guglielmo da Saliceto: nuove ricognizioni sulla tradizione in volgare'. *Filologia Italiana* 5 (2008): 53–101.

[6] De Angelis, Alessandro. 'Due canti d'amore in grafia greca dal Salento medievale e alcune glosse greco-romanze'. *Cultura* LXX, no. 3/4 (2010): 371–413.

[7] Den Heijer, Johannes, Andrea Schmidt, e Tamara Pataridze. *Scripts beyond Borders: A Survey of Allographic Traditions in the Euro-Mediterranean World*. Vol. 62. Institut Orientaliste. Peeters: Louvain-la-Neuve, 2014.

[8] Maggiore, Marco. 'Sui testi romanzi medievali in grafia greca come fonte di informazione linguistica'. *Zeitschrift Für Philologie* 133, no. 2 (2017): 313–342.

[9] Maggiore, Marco, e Daniele Arnesano. 'La formula matrimoniale del Codice Hunter 475: il testo più antico in volgare siciliano?' *Bollettino del Centro di studi filologici e linguistici siciliani* XXXI (2020).

[10] Proverbio Vania, Delio. ‘An Italian Text in Syro-Xenic Clothes: The Italo-Garšūnī Pasquin Borg. Ar. 278, Ff.1r-2v’. *Rivista Degli Orientali* 93, no. 1 (2020): 93–107.

[11] Rubin, Aaron David. ‘Judeo-Italian’. In *Handbook of Jewish Languages*, edited by Lily Kahan and Aaron David Rubin. Leida: Brill, 2016.

The dark mirror of artificial intelligence: how AI affects climate change

Mauro De Bari
University of Bari Aldo Moro, Italy – mauro.debari@uniba.it

ABSTRACT

In the aftermath of the conclusion of the Digital Transformation (DT) Era, society is undergoing a pivotal shift, notably marked by the widespread integration of Artificial Intelligence (AI). This paper explores the multifaceted impact of AI on global climate change, focusing on the GLAM (galleries, libraries, archives, and museums) sector and the Cultural and Creative Industries (CCIs) and emphasising the European approach to address these challenges. As AI permeates various sectors, especially museums utilising AI for enhanced experiences and heritage preservation, the escalating computational demands contribute significantly to carbon emissions, demanding urgent intervention. Europe, exemplified by initiatives like the European Green Deal, underscores a commitment to sustainable development in mitigating environmental repercussions.

This research delves into the moral and cultural dimensions of the AI-climate change nexus caused by European cultural institutions, contributing to the ongoing discourse on responsible AI development. A case study on OpenAI chatbot ChatGPT highlights the imperative to educate users about responsible AI usage, preventing irreversible damage to the environment and communities. Overall, the author seeks a harmonious balance between technological innovation and ethical responsibility in navigating the complexities of the AI-driven Era.

KEYWORDS

Digital; Artificial Intelligence (AI); climate change; cultural implications; ChatGPT.

1. INTRODUCTION

Nowadays, at the end of the bridge period called the Digital Transformation (DT) Era, communities are living in a new transitional path [12]. To reinforce this concept, Nuspire's chief security officer, J.R. Cunningham[1], recently argued why this crucial conclusion assumes natural change and its implications for society in the by-now digital world. In particular, the spread of Artificial Intelligence (AI) represents the new main character and one of the "natural causes" of this epilogue, becoming the fulcrum of scientific and academic debates [13].

The widespread integration of AI technologies in contemporary society has ushered in a new era of innovation, transforming the GLAM sector cultural institutions and industries, especially the Cultural and Creative Industries (CCIs), reshaping how we navigate the modern world's intricacies. However, this remarkable advancement in AI adoption brings forth many challenges, prominently the escalating impact on global climate change. This paper explores the nuanced dimensions of AI's negative influence on climate change, specifically focusing on the ethical and cultural ramifications. The first European proposals to address this issue will be argued, highlighting the distinctive cultural approach to addressing the complex interplay between AI, cultural institutions, and climate change.

AI has effectively affected the entire cultural sector, especially museums, which have invested in AI technologies to create engaging user experiences and preserve heritage more efficiently.

However, as AI systems continue to increase exponentially, the accompanying surge in computational demands, data storage, and energy consumption substantially contributes to the escalating carbon emissions driving the global climate crisis. The scale of this environmental impact, with direct and indirect repercussions [14], necessitates urgent scrutiny and strategic intervention to mitigate potentially irreversible consequences. The European approach to this challenge is characterised by an increasing commitment to sustainable development, as evidenced by initiatives such as the European Green Deal[2], which aims to make the European Union climate-neutral by 2050; "the transition to a climate-neutral society is both an urgent challenge and an opportunity to build a better future for all"[3].

Beyond the ecological sphere, the intricate relationship between AI and climate change extends to the core of societies, posing global challenges to cultural ecosystems. Traditional ways of life are disrupted, and communities grapple with the

[1] https://enterprisersproject.com/user/jr-cunningham
[2] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_it
[3] https://climate.ec.europa.eu/eu-action/climate-strategies-targets/2050-long-term-strategy_en

repercussions of a changing climate, often exacerbated by AI-driven practices [6]. The European perspective emphasises the importance of preserving cultural heritage and values in the face of technological advancements. For this reason, initiatives like the European AI Alliance and the European Commission's Ethics Guidelines for Trustworthy AI[4] underscore a commitment to aligning AI development with European values, ensuring a balance between innovation and ethical responsibility.

Recently, the European Parliament[5] and Council[6], and as a consequence, Europe, has set the stage for a cutting-edge change in the regulation of artificial intelligence. The aim is to protect and safeguard European citizens, classifying AI systems "as high-risk (due to their significant potential harm to health, safety, fundamental rights, environment, democracy and the rule of law), clear obligations were agreed"[7]. As written above, this article seeks to reflect on the moral and cultural dimensions of the nexus between climate change and artificial intelligence in the European context, addressing a topic highly discussed but lacking scientific literature that ponders on the pollution caused by the excessive use of AI in the cultural sector, especially the museum sector. This contributes to the ongoing debate on the responsible development of AI, in line with the European commitment to sustainability and cultural preservation. The analysis examines the possible harmful impact of AI on climate change. It supports a sustainable and inclusive future in which cultural heritage and sustainable cultural progress are integral components of the evolving narrative on AI and its environmental repercussions. Notably, the author reports a case study focusing on one of the most widespread AI technologies, ChatGPT, demonstrating that it is necessary to educate users to wisely use the potential offered by OpenAI in order not to commit irreversible damage to the environment and communities.

2. AI IN CULTURAL SYSTEM AND HOW AFFECTS IT: THE CLIMATE CHANGE ISSUE

Integrating Artificial Intelligence (AI) within the cultural system, especially regarding climate change, necessitates transformative shifts that influence environmental dynamics, institutional practices, and collective behaviours in a rapidly evolving socio-ecological landscape [17]. AI technologies enable cultural institutions like museums, galleries, and heritage sites to develop innovative solutions for assessing and mitigating their ecological footprint and environmental impact. Using machine learning algorithms, data analytics platforms, and IoT sensors, these institutions can predict energy consumption and resource management, fostering sustainability in the context of climate change mitigation and adaptation [9].

However, integrating AI tools raises several environmental concerns due to the energy consumption and electronic waste generated by AI-based systems within cultural institutions. While AI is a vanguard of innovation, its expansive carbon footprint is increasingly alarming. Central to this concern is the escalating energy consumption inherent in training sophisticated AI models. Disturbingly, since 2012, the computational demands for such models have escalated, doubling every 3.4 months, portending a significant uptick in greenhouse gas emissions. Projections underscore this urgency, with estimates indicating that 2040 emissions from the Information and Communications Technology (ICT) sector could burgeon to constitute 14% of global emissions [8].

This contributes to carbon emissions and ecological degradation [7], necessitating environmentally responsible approaches to minimise these impacts [12]. Additionally, the supply chains and waste management practices associated with AI hardware components exacerbate sustainability challenges globally. Thus, stakeholders must adopt holistic and ethical strategies to mitigate the adverse environmental effects while leveraging AI's potential for innovation.

Socially, AI's integration within cultural institutions [1] presents ethical and socio-political implications related to accessibility, inclusivity, and democratic participation. As AI-driven technologies like VR, AR, and algorithmic curation systems enhance visitor experiences and educational outreach, questions arise about equitable access, representation, and empowerment within the cultural ecosystem. Critical engagement and socio-cultural awareness are essential to navigate the complexities between technological innovation, cultural heritage, and social justice, promoting inclusive dialogues on climate change and planetary well-being.

[4] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[5] https://www.europarl.europa.eu/portal/en
[6] https://www.consilium.europa.eu/it/european-%20council/
[7] European Parliament. 2023. "Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI." News European Parliament. December 09. https://www.europarl.europa.eu/news/en/pressroom/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.

Museums play a crucial role in this landscape [16], evolving their practices by leveraging AI technologies to reimagine curatorial approaches, educational initiatives, and community outreach programs. AI-enabled immersive experiences further enhance public engagement, cultivating empathy and advocacy for environmental sustainability.
However, ethical considerations remain paramount. Museums must critically reflect on their practices, partnerships, and representational strategies to ensure inclusivity, respect for diverse voices, and recognition of indigenous knowledge systems[8]. By adopting this approach, they can foster social justice and diversity within the cultural ecosystem.
Finally, integrating AI within museums is a complex endeavour that intersects with environmental, social, and cultural aspects, shaping their role in addressing climate change and planetary well-being. Through interdisciplinary collaboration, ethical reflexivity, and transformative innovation, museums can harness AI's potential to realise their mission of fostering resilience, sustainability, and social justice in a globally interconnected context.

3. IN THE CHATGPT CASE, POPULAR DOES NOT MEAN GOOD

As previously reported, AI systems receive a lot of visibility and credibility in contemporary society. One of the most popular AI tools is the OpenAI company's Chat Generative Pre-trained Transformer (ChatGPT)[9] product. However, in this exponential increase, the immoderate utilisation by users is taking over [4]. In particular, the cultural framework's overuse of ChatGPT raises significant inquiries about the ethical consequences of AI technology in safeguarding, interpreting, and propagating cultural heritage, mainly when it is employed unnecessarily or inappropriately. Some potential ethical concerns that arise when using ChatGPT in the cultural system include algorithmic bias, data privacy, intellectual property rights, and a human-centred approach to cultural preservation, interpretation, and dissemination [10]. Not least, climate change occupies a prominent position in the framework of the pros and cons of using and abusing the ChatGPT AI system [2]. For this reason, there is an urgent need to establish transparent, accountable and participatory frameworks that prioritise ethical integrity, cultural sensitivity and social responsibility in harnessing the transformative potential of ChatGPT within the cultural landscape. Furthermore, museums and cultural institutions must ensure responsible implementation of AI when using ChatGPT by critically reflecting on the ethical dimensions of AI-mediated interactions, decision-making processes and knowledge production within the cultural ecosystem[10].
Furthermore, the deployment of ChatGPT within the cultural landscape facilitates the preservation, interpretation, and democratisation of cultural heritage, knowledge, and artistic expressions across various mediums and disciplines. By incorporating ChatGPT into digital repositories, online exhibitions, and interactive installations, cultural institutions can curate dynamic and interactive experiences that transcend geographical boundaries, temporal constraints, and linguistic barriers [3]. ChatGPT's multilingual capabilities and adaptability empower museums, galleries, and cultural organisations to reach broader audiences, facilitate cross-cultural dialogues, and promote intercultural understanding by translating, contextualising, and interpreting diverse cultural narratives, artefacts, and perspectives. Additionally, ChatGPT's capacity to generate content, curate collections, and facilitate collaborative engagements enables cultural institutions to reimagine their curatorial practices, educational initiatives, and community outreach programs, thereby revitalising cultural landscapes, fostering global connections, and preserving intangible cultural heritage for future generations.
Moreover, an emergent concern lies in the unintentional environmental ramifications engendered by museums' adoption of ChatGPT and similar AI infrastructures. Despite ChatGPT's intrinsic text-based nature ostensibly eschewing direct greenhouse gas emissions, the requisite computational prowess for its training, deployment, and maintenance precipitates significant energy expenditures. ChatGPT consumes between 10 and 100 times more energy than e-mail. According to research from the University of Washington [5], the energy used for hundreds of millions of ChatGPT queries could be equivalent to that consumed by 33,000 U.S. households in a single day, as reported by Yahoo! Finance. Sajjad Moazeni, a professor of electrical and computer engineering at UW, informed Yahoo! Finance that a ChatGPT query likely requires "10 to 100 times more energy" than e-mailing. This energy-intensive modus operandi, predominantly sustained by fossil fuel-reliant infrastructures, amplifies carbon emissions, augments electronic waste generation, and exacerbates resource depletion [11].
It is estimated that ChatGPT may consume up to 1/2 litre of water to handle between 5 and 50 queries, thus contributing to significant water consumption, which varies depending on the season and the server's location. Companies like Microsoft and Google have reported substantial increases in water consumption associated with using Artificial Intelligence, with

[8] https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
[9] https://chat.openai.com/auth/login
[10] https://cuseum.com/blog/2023/4/13/9-ways-chatgpt-can-empower-museums-cultural-organizations-in-the-digital-age

gains of +34% and +30%, respectively, between 2021 and 2022. To mitigate these impacts, it is essential to place servers in climatically favourable areas like Iowa to leverage more efficient cooling conditions[11].

Consequently, an underexplored academic analysis underscores the burgeoning pollution footprint inadvertently propagated by museums integrating ChatGPT. The overarching proliferation of such AI technologies across cultural landscapes accentuates strains on global energy infrastructures, thereby engendering cascading ecological repercussions. In an academic *milieu*, it becomes imperative to interrogate, elucidate, and proffer mitigative strategies addressing this nascent yet consequential nexus between cultural institutions' AI adoption, environmental degradation, and sustainability imperatives. Thus, scholarly endeavours must concurrently navigate the confluence of technological innovation, ethical integrity, and environmental stewardship to forge a sustainable trajectory for cultural heritage conservation in the digital epoch.

4. CONCLUSION

AI development and utilisation often operate within a "shadowy sphere", fostering a lack of transparency and accountability regarding environmental repercussions. Some companies prioritise their financial gains and competitive advantages over acknowledging and addressing AI technologies' potential adverse environmental impacts. This opacity makes it challenging for users to fully understand the ecological footprint left by AI systems, as the intricate nature of these technologies obscures accurate assessments of their carbon emissions or broader environmental effects, hindered further by the clandestine methods and concealed data utilised in AI model training. Addressing this issue necessitates the establishment of more transparent protocols and regulations that align AI development and deployment with environmental considerations. Adopting a responsible approach to AI that emphasises sustainability requires enhanced accountability mechanisms.

ChatGPT is a significant milestone in computational linguistics and artificial intelligence in today's rapidly evolving technological landscape, offering multifaceted benefits across various sectors. However, it is crucial to contextualise this technological leap within a comprehensive environmental and ethical framework, especially concerning human-induced climate change. While ChatGPT may not directly contribute to environmental degradation, its operational framework relies heavily on energy-intensive data centres and computational resources, leading to substantial carbon footprints that could aggravate ecological disruptions. Therefore, the deployment of ChatGPT necessitates carefully considering epistemological and ethical concerns. Its responsible utilisation demands robust regulatory oversight, sustainable infrastructure development, and moral conduct.

As custodians of this emerging technology, stakeholders must navigate the intricate intersection between technological innovation and environmental stewardship with utmost diligence. This journey involves a multifaceted strategy encompassing sustainability assessments, energy-efficient computing models, and ethical guidelines prioritising ecological sustainability, social equity, and long-term environmental well-being. While ChatGPT represents a pinnacle of technological innovation and computational capability, its effective integration into societal structures hinges on an unwavering commitment to ethical responsibility and environmental sustainability. As scholars, practitioners, and policymakers traverse the intricate landscape of AI-driven progress, an uncompromising focus on responsible innovation, ethical introspection, and sustainable practices remains crucial. Ultimately, the impact of ChatGPT on climate change mitigation hinges on our collective determination to harness its capabilities judiciously, forging a harmonious synergy between technological progress and global ecological health.

REFERENCES

[1] Avik Sinha, Arnab Adhikari, e Ashish Kumar Jha. «Innovational duality and sustainable development: finding optima amidst socio-ecological policy trade-off in post-COVID-19 era». *Journal of Enterprise Information Management*, fasc. February 18. (2022).

[2] Bassetti, Francesco. «ChatGPT: climate knowledge – and misinformation – at your fingertips». *Foresight*, 24 agosto 2023. https://www.climateforesight.eu/interview/chatgpt-climate-knowledge/

[3] Charr, Manuel. «Museum Uses Artificial Intelligence to Curate Better Exhibitions». *MuseumNext*, 23 giugno 2021. https://www.museumnext.com/article/museum-uses-artificial-intelligence-to-curate- better-exhibitions/

[4] Chow, Andrew R. «How ChatGPT Managed to Grow Faster Than TikTok or Instagram». *TIME*, 8 febbraio 2023. https://time.com/6253615/chatgpt-fastest-growing/

[5] Cohan, Peter. «As ChatGPT And Other AI Tools Increase Energy Demand, Here's What Investors Need To Know». *Forbes*, 9 novembre 2023. https://www.forbes.com/sites/petercohan/2023/11/09/equinix-and-vertiv- stock-prices-could-rise-on-generative-ais-energy-use/?sh=45c53bae6685

[11] https://iiai.uiowa.edu

[6] Coleman, Jude. «AI's Climate Impact Goes Beyond Its Emissions. To understand how AI is contributing to climate change, look at the way it's being used». *SCIAM*, 7 dicembre 2023. https://www.scientificamerican.com/article/ais-climate-impact-goes- beyond-its- emissions/#:~:text=But%20as%20AI%27s%20popularity%20keeps,way%20AI%20affects%20the%20climate

[7] De Vries, Alex. «The growing energy footprint of artificial intelligence». *Joule* 7, fasc. 10 (2023): 2191–2194.

[8] Kanugo, Alokya. «The Green Dilemma: Can AI Fulfil Its Potential Without Harming the Environment?» *Earth.org*, 2023. https://earth.org/the-green-dilemma-can-ai-fulfil-its-potential-without-harming-the-environment/

[9] Lee, Da-sheng, Yang-Tang Chen, e Shih-Lung Chao. «Universal workflow of artificial intelligence for energy saving». *Energy Reports* 8 (novembre 2022): 1602–1633.

[10] Mouriquand, David. «ChatGPT turns one: The birthday threatening culture and creativity». *Euronews.culture*, novembre 2023. https://www.euronews.com/culture/2023/11/30/chatgpt-turns-one-the-birthday-threatening- culture-and- creativity

[11] O'Brien, Matt, e Hannah Fingherut. «Artificial intelligence technology behind ChatGPT was built in Iowa — with a lot of water». *AP*, 9 settembre 2023. https://apnews.com/article/chatgpt- gpt4-iowa-ai-water- consumption-microsoft-f551fde98083d17a7e8d904f8be822c4

[12] Panetta, Kasey. «Keep AI From Doing More Climate Harm Than Good». *Gartner*, 28 agosto 2023. https://www.gartner.com/en/articles/keep-ai-from-doing-more-climate-harm-than-good.

[13] Poremba, Sue. «What comes after the digital transformation?» *Security Intelligence*, 25 aprile 2023.

[14] Sissa, Giovanna. «Intelligenza artificiale e cambiamenti climatici: rischi e opportunità». *Agenda Digitale*, 26 settembre 2019. https://www.agendadigitale.eu/cultura-digitale/intelligenza-artificiale- e-cambiamenti-climatici- rischi-e-opportunita/

[15] Styx, Lauren. «How are museums using artificial intelligence, and is AI the future of museums?» *MuseumNext*, 18 giugno 2023. https://www.museumnext.com/article/artificial-intelligence- and-the-future-of-museums/

[16] Werner, John. «Museum Curation In The Age Of AI». *Forbes*, 8 gennaio 2024. https://www.forbes.com/sites/johnwerner/2024/01/08/museum-curation-in-the-age-of-ai/?sh=773185556f98

[17] Zhao, Jingchen, e Beatriz Gomez Farinas. «Artificial Intelligence and Sustainable Decision». *European Business Organization Law Review* 24, fasc. 1 (2023): 1–39.

Un sistema di classificazione automatica di immagini relative a materiali librari antichi e moderni

Nicola Barbuti[1], Tommaso Caldarola[2]
[1] Università degli Studi di Bari Aldo Moro, Italia - nicola.barbuti@uniba.it
[2] D.A.BI.MUS. S.r.l., Italia - tommaso@caldarola.net

ABSTRACT[1]

Nell'ambito della digitizzazione dei beni librari, tra le sfide tecniche più significative a oggi ancora irrisolte vi è la classificazione automatica delle immagini, processo che combina l'informatica, la biblioteconomia e le tecnologie dell'informazione per categorizzare e organizzare digitalmente le strutture degli oggetti digitali che riproducono volumi antichi manoscritti e a stampa. Questo articolo presenta una recente ricerca sperimentale di tre modelli di classificazione automatica di immagini digitali che riproducono manoscritti e libri antichi e moderni, finalizzata a estrarre dal layout le informazioni relative alla struttura dei volumi per la codifica nei metadati di gestione delle immagini.

PAROLE CHIAVE

Classificazione automatica; nomenclature; digitizzazione; apprendimento profondo; reti neurali convoluzionali CCN.

1. INTRODUZIONE

Sebbene la classificazione automatica sia uno dei settori di ricerca di più antica tradizione nell'ambito delle digital humanities [1][2], nei processi di metadatazione e gestione delle immagini di materiali biblioteconomici (libri antichi, manoscritti, periodici, ecc.) essa non ha mai trovato applicazione, a causa della struttura estremamente complessa degli originali rappresentati e delle caratteristiche formali del layout degli oggetti digitali. Quasi mai, infatti, le immagini esposte in rete relative alle diverse parti di cui si compone la struttura di un volume, soprattutto antico, riportano la *nomenclature* dell'elemento nella visualizzazione, come da regola prevista nelle linee guida dell'ICCU[3]. Questo perché, a oggi, la sola opzione per valorizzare le informazioni di struttura nei metadati che gestiscono ciascuna immagine consiste nell'inserire direttamente nei filename la *nomenclature* della parte rappresentata.

Tale opzione, però, è difficilmente praticabile, in quanto richiede un ingente lavoro manuale di inserimento delle *nomenclature* nei filename o preliminarmente alla scansione delle diverse parti, di modo da ritrovarle poi nei metadati dell'immagine, oppure alla fine del ciclo di scansione dell'intero volume, modificando i filename di tutte le immagini tramite appositi tool di rinomina. Entrambe le soluzioni presentano criticità rilevanti, in quanto la necessità di modificare il filename più volte durante la scansione ne rallenta sensibilmente il ritmo e, nel contempo, aumenta il rischio di errori che generano poi ripercussioni sulle altre fasi di digitizzazione; la rinomina di tutti i file, assegnando automaticamente le differenti *nomenclature* a insiemi di immagini che rappresentano le varie parti della struttura del volume, differisce da libro a libro.

Data l'elevata variabilità delle strutture degli originali, soprattutto se antichi, generare un unico modello di classificazione automatica sempre valido per tutte le tipologie di originali digitizzati è una sfida scientifica di alto livello.

Da queste premesse muove la ricerca che si presenta, finalizzata a sviluppare e testare modelli avanzati di classificazione automatica dei layout di immagini digitali riproducenti manoscritti, libri antichi e libri moderni, che siano in grado di identificare, estrarre e codificare le informazioni sulle strutture. L'approccio adottato per la progettazione e lo sviluppo è stato basato sulle più recenti tecniche di *apprendimento profondo* basate su *reti neurali profonde*. In particolare, sono state utilizzate *reti neurali convoluzionali* (CNN) [2][4], nell'ottica di sfruttare quanto più possibile le caratteristiche distintive rilevabili da immagini digitizzate che rappresentano manoscritti, libri antichi e libri moderni.

[1] Nicola Barbuti ha curato il par. 1, parte del par. 3, Bibliografia e revisione finale; Tommaso Caldarola ha curato i par. 2 e 3, Bibliografia e revisione finale.
[2] Non si registrano riferimenti bibliografici a studi scientifici recenti sulla classificazione automatica di immagini relative al digital heritage, topic oggetto del presente lavoro. Si rinvia, pertanto, a studi relativi ad alcune sperimentazioni di classificazione che hanno attinenza con quanto attuato nella ricerca.
[3] https://www.internetculturale.it/getFile.php?id=44402
[4] https://it.wikipedia.org/wiki/Rete_neurale_convoluzionale

Nel processo di progettazione e addestramento del modello sono state utilizzate *TensorFlow*, una piattaforma di machine learning end-to-end[5], e la libreria API ad alto livello *Keras*[6] che ne facilita l'utilizzo [4]. Sono state inoltre utilizzate librerie di manipolazione delle immagini e librerie per operazioni matematiche su array.
La metodologia di sperimentazione ha previsto la raccolta di un dataset di immagini rappresentativo, la preparazione accurata dei dati, la predisposizione di un set di almeno 20 esempi di *nomenclature* di ciascuna tipologia di libro per l'addestramento dei modelli di classificazione e la validazione degli esiti dell'apprendimento, la definizione di un'architettura di reti neurali per catturare le specificità delle categorie target.
Infine, a valle dei test applicativi, è stata valutata l'efficacia del modello tramite una serie di metriche di prestazione, includendo l'accuratezza nella classificazione delle immagini, la capacità di generalizzazione su nuovi dati e l'interpretabilità delle decisioni del modello.

1. METODOLOGIA

Preparazione dei dati per l'addestramento

Per la preparazione dei dati per l'addestramento sono stati selezionati set di immagini campione relative alle varie parti delle strutture delle tre tipologie di libri digitizzati da riconoscere e classificare. Dal formato originario, i campioni sono stati ridimensionati a 224x224 pixel senza mantenere l'*aspect ratio*. Di questo set, l'80% è stato scelto in maniera casuale e impiegato per l'addestramento, il restante 20% è stato utilizzato per la validazione del modello. L'ottimizzazione delle prestazioni durante l'addestramento dei modelli e la gestione dei dati sono stati aspetti cruciali per l'efficacia della sperimentazione. Sono stati rispettati i tre concetti fondamentali relativi alla gestione dei dati:

1. *cache*: caricare i dati in memoria prima di iniziare l'addestramento ha consentito di ridurre il tempo di caricamento durante ogni fase, migliorando le prestazioni complessive;
2. *shuffle*: la mescolanza casuale degli esempi di addestramento è stata utile a evitare che il modello imparasse dai dati sequenze indesiderate;
3. *prefetch*: la possibilità di caricare i dati per l'esecuzione della fase successiva mentre il modello è ancora in fase di addestramento ha sensibilmente ridotto i tempi di attesa durante il caricamento.

L'utilizzo appropriato di queste tecniche ha migliorato le prestazioni, riducendo i tempi di attesa.

Normalizzazione

Durante l'addestramento, la tecnica di *normalizzazione dei dati* è indispensabile per standardizzare le caratteristiche (*features*) dei dataset in modo che abbiano una scala comune, rendendo più agevole l'apprendimento dei pesi associati a ciascuna caratteristica. Senza normalizzazione, *features* con valori più grandi avrebbero potuto dominare, influenzando negativamente l'addestramento e compromettendo il corretto apprendimento. Per la sperimentazione dei modelli, si è scelto di adottare la *Batch Normalization*, una tecnica impiegata nell'addestramento delle reti neurali per migliorare la stabilità e l'accelerazione della convergenza. In pratica, la *Batch Normalization* normalizza l'output di ogni strato nascosto (o input dello strato) considerando il batch di dati su cui viene eseguito. L'obiettivo è rendere l'output del layer più stabile durante l'addestramento. Questa tecnica è particolarmente utile quando si utilizzano *reti neurali profonde*.
L'utilizzo della *Batch Normalization* ha prodotto tassi di apprendimento più elevati, una regolarizzazione del processo e la riduzione del rischio di *vanishing/exploding gradients*.

Sviluppo dei modelli

I modelli sono costituiti da CNN complesse. In particolare, sono implementazioni di *MobileNetV2* [3][7], una variante di *MobileNet* progettata per essere efficiente in termini computazionali e adatta per l'esecuzione su dispositivi anche con risorse limitate, come ad esempio i dispositivi mobili.
Alcuni punti chiave di ciascun modello sono:

1. Struttura Generale: il modello ha una struttura gerarchica con diversi blocchi, ciascuno contenente strati convoluzionali profondi. I blocchi seguono una struttura di base comune, che comprende strati di espansione, strati di profondità-wise separable convolution, e strati di proiezione;
2. Depthwise Separable Convolution: molte delle convoluzioni nel modello sono di tipo "depthwise separable"; queste parti separano la convoluzione in due fasi: una convoluzione separata per ciascun canale (*depthwise*) seguita da una convoluzione 1x1 (*pointwise*) per mescolare le informazioni spaziali e di canale;

[5] https://www.tensorflow.org/?hl=it.
[6] https://keras.io/.
[7] https://it.mathworks.com/help/deeplearning/ref/mobilenetv2.html.

3. Bottleneck Blocks: alcuni blocchi sono stati progettati come blocchi *bottleneck*, riducendo la dimensionalità nei primi strati e quindi espandendola nuovamente;

4. Dimensioni Progressive: le dimensioni delle *feature map* sembrano diminuire progressivamente attraverso i blocchi, con i blocchi più profondi che gestiscono feature map più piccole;5. Ultimo Strato: l'ultimo strato (`Conv_1`) produce una feature map con 1280 canali, seguito da un livello di Batch Normalization e un'attivazione ReLU ('out_relu')[8];

6. Dimensioni dell'Input: l'input atteso è un tensore di forma (224, 224, 3), che suggerisce immagini RGB di dimensione 224x224 pixel.

Ogni modello è diviso in due *parti sequenziali*[9]. La prima è complessa ed è costituita da una struttura di tipo funzionale (*Functional*), mentre la seconda parte è costituita da due strati densi. Nello specifico, essendo i modelli da sviluppare relativi ad artefatti molto simili tra loro, la prima parte sequenziale è costituita da uno strato molto complesso che prevede 16 blocchi costituiti da una determinata struttura.

Per rendere l'idea della complessità di ciascun blocco all'interno di una rete neurale convoluzionale, si riporta di seguito il dettaglio di uno:

1. block_1_expand (Conv2D): questo strato è un'operazione di convoluzione 2D (Conv2D) denominata "block_1_expand" con una forma di output (None, 112, 112, 48). La convoluzione ha 384 parametri e riceve l'input da un livello chiamato 'expanded_conv_project_BN[0][0]';

2.block_1_expand_BN (BatchNormalization): questo strato esegue la normalizzazione batch (Batch Normalization) sull'output del livello di convoluzione precedente ('block_1_expand[0][0]');

3. block_1_expand_relu (ReLU): questo strato applica l'attivazione ReLU (Rectified Linear Unit) all'output del livello di normalizzazione batch precedente ('block_1_expand_BN[0][0]');

4. block_1_pad (ZeroPadding2D): questo strato aggiunge zero padding all'output del livello ReLU ('block_1_expand_relu[0][0]');

5. block_1_depthwise (DepthwiseConv2D): questo è uno strato di convoluzione profonda (depthwise convolution) denominato "block_1_depthwise" con una forma di output (None, 56, 56, 48). Ha 432 parametri e riceve l'input dal livello 'block_1_pad[0][0]';

6. block_1_depthwise_BN (BatchNormalization): questo strato esegue la normalizzazione batch sull'output del livello di convoluzione profonda ('block_1_depthwise[0][0]');

7. block_1_depthwise_relu (ReLU): applica l'attivazione ReLU all'output del livello di normalizzazione batch precedente ('block_1_depthwise_BN[0][0]');

8. block_1_project (Conv2D): questo strato esegue un'ulteriore convoluzione 2D chiamata "block_1_project" con una forma di output (None, 56, 56, 8). Ha 384 parametri e riceve l'input dal livello 'block_1_depthwise_relu[0][0]'.

9. block_1_project_BN (BatchNormalization): esegue la normalizzazione batch sull'output del livello di convoluzione precedente ('block_1_project[0][0]').

La seconda parte sequenziale ha due strati *densi*, che sono una tipologia tra le più comuni nelle reti neurali artificiali. Uno strato denso è caratterizzato da tre componenti principali:

1. Connessioni completamente collegate: ogni neurone in uno strato denso è connesso a ciascun neurone nello strato successivo. Questo significa che c'è una connessione diretta tra ogni coppia di neuroni;

2. Pesi: ogni connessione tra neuroni ha un peso associato. Questi pesi vengono appresi durante il processo di addestramento della rete neurale;

3. Funzione di attivazione: ogni neurone ha una funzione di attivazione associata che determina l'output del neurone dati i suoi input pesati. La funzione di attivazione introduce non linearità nella rete.

Addestramento

Quando si addestra una Rete Neurale Convoluzionale (CNN) come *MobileNetV2* per la classificazione di immagini, il processo di addestramento si concentra sulla capacità della rete di estrarre automaticamente le caratteristiche rilevanti

[8] Un'attivazione ReLU indica che l'ultimo strato del modello, denominato "out_relu", utilizza una funzione di attivazione ReLU (Rectified Linear Unit), ampiamente utilizzata nelle reti neurali, soprattutto in contesti di reti convoluzionali. La funzione ReLU definisce la sua uscita come zero per tutti gli input negativi e lineare per gli input positivi. Matematicamente, è espressa come: f(x) = max(0, x), dove x è l'input della funzione. Quindi, se l'input è positivo, la funzione restituirà l'input stesso; se l'input è negativo, la funzione restituirà zero.

[9] In un contesto di reti neurali, una parte sequenziale si riferisce a un tipo specifico di architettura modello. A esempio, in *Keras* è chiamata "Sequential Model". In un modello sequenziale, gli strati vengono aggiunti uno dopo l'altro in sequenza. Ogni strato ha un solo input e un solo output, creando così una sequenza lineare di strati. Questa è un'architettura molto comune, particolarmente adatta per la costruzione di reti neurali dette *feedforward*.

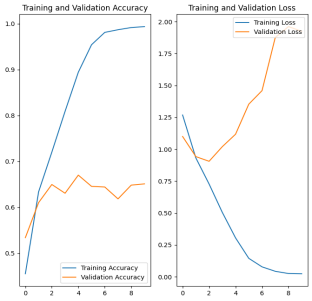
dall'immagine per discriminare tra diverse classi. Durante il processo, le prime fasi eseguono una serie di operazioni di *convoluzione* e *pooling* per estrarre feature di basso livello come bordi, texture e pattern semplici dall'immagine in ingresso. Man mano che l'informazione attraversa la rete, le feature estratte diventano via via più complesse e astratte, rappresentando concetti di livello superiore come forme, parti di oggetti e strutture. Questo avviene grazie alla profondità della rete neurale, che permette la combinazione di feature di basso livello per creare rappresentazioni sempre più sofisticate dell'immagine. La rete, grazie alla sua struttura stratificata, è in grado di catturare informazioni contestuali e relazioni spaziali tra le feature estratte. Questo processo le permette di comprendere meglio il contesto in cui si trovano le feature e di utilizzare queste informazioni per migliorare la discriminazione tra classi. Una volta estratte e processate le feature dall'immagine, la rete CNN passa attraverso uno o più strati completamente connessi che convertono le feature estratte in una distribuzione di probabilità sulle classi di output. In questo modo, durante l'addestramento, le reti neurali convoluzionali imparano a riconoscere automaticamente le feature rilevanti dall'immagine tramite un processo di estrazione e stratificazione delle feature, e poi utilizzano queste informazioni per effettuare predizioni accurate sulla classe dell'immagine in ingresso.
La fase di addestramento del sistema ha previsto il passaggio dei dati per l'apprendimento attraverso ciascun modello tramite un processo di *forward propagation*. In questo modo è stato possibile calcolare la perdita tra le previsioni del modello e i valori desiderati. Allo scopo, sono stati utilizzati la funzione di *retropropagazione* per calcolare i gradienti della perdita rispetto ai pesi del modello e un ottimizzatore per aggiornarli e ridurre la perdita.
Il processo di addestramento tipicamente si articola in *epoche*, dove un'epoca rappresenta un passaggio completo attraverso l'intero set di dati utilizzato. Nel caso specifico, ne sono state utilizzate dieci relative ai dati di addestramento e ai dati di validazione.
Durante l'addestramento, ciascun modello ha utilizzato il set di immagini di riferimento per regolare i pesi in base alla funzione di perdita definita. Il set di validazione è stato utilizzato per monitorare le prestazioni del modello applicato a dati non impiegati nell'addestramento e valutare l'eventuale sovraccarico. Il risultato dell'addestramento con le informazioni delle perdite e delle metriche durante ogni epoca è stato memorizzato, nell'ottica di utilizzarlo successivamente per tracciare la curva di apprendimento e valutare le prestazioni di ciascun modello. Di seguito è rappresentato un log di addestramento durante l'allenamento di una rete neurale utilizzando *TensorFlow* e *Keras* (vd. Fig. 1).

	A
1	Epoch 1/10
2	12s 164ms/step - loss: 1.8127 - accuracy: 0.4573 - val_loss: 0.9502 - val_accuracy: 0.7082
3	Epoch 2/10
4	9s 150ms/step - loss: 1.0249 - accuracy: 0.6368 - val_loss: 0.8574 - val_accuracy: 0.7339
5	Epoch 3/10
6	9s 146ms/step - loss: 0.7904 - accuracy: 0.7147 - val_loss: 0.7265 - val_accuracy: 0.7296
7	Epoch 4/10
8	9s 146ms/step - loss: 0.6143 - accuracy: 0.7639 - val_loss: 0.7797 - val_accuracy: 0.6953
9	Epoch 5/10
10	9s 160ms/step - loss: 0.5589 - accuracy: 0.7735 - val_loss: 0.7101 - val_accuracy: 0.7210
11	Epoch 6/10
12	9s 146ms/step - loss: 0.4307 - accuracy: 0.8312 - val_loss: 0.6260 - val_accuracy: 0.7039
13	Epoch 7/10
14	9s 148ms/step - loss: 0.4247 - accuracy: 0.8280 - val_loss: 0.6795 - val_accuracy: 0.7768
15	Epoch 8/10
16	9s 149ms/step - loss: 0.3511 - accuracy: 0.8483 - val_loss: 0.6079 - val_accuracy: 0.7597
17	Epoch 9/10
18	9s 153ms/step - loss: 0.3399 - accuracy: 0.8579 - val_loss: 0.8017 - val_accuracy: 0.6824
19	Epoch 10/10
20	9s 147ms/step - loss: 0.3447 - accuracy: 0.8408 - val_loss: 0.7175 - val_accuracy: 0.7124

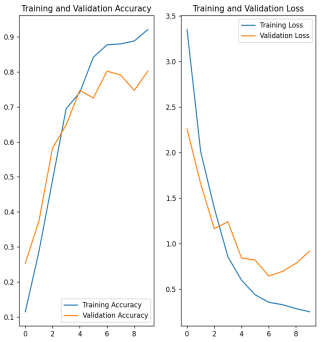
Figura 1. Log di addestramento durante l'allenamento di una rete neurale

- Epoch 1/10: indica che il modello sta attraversando la prima epoca di addestramento;
- 12s: rappresenta il tempo impiegato per completare l'epoca corrente;
- 164ms/step: indica il tempo medio impiegato per processare ogni passo (batch) durante l'epoca. Un passo è un singolo aggiornamento dei pesi del modello, eseguito su un batch di dati. Un valore basso è generalmente desiderato, poiché indica un addestramento più veloce;
- loss: è il valore della funzione di perdita sul set di dati di addestramento. La funzione di perdita misura quanto il modello si discosta dalla verità rispetto alle sue previsioni durante l'addestramento. L'obiettivo è ridurre questo valore;
- accuracy: rappresenta l'accuratezza del modello sul set di dati di addestramento, misurata come la percentuale di previsioni corrette. In questo caso, l'accuratezza è del 45,73%;
- val_loss: è il valore della funzione di perdita sul set di dati di validazione. Il set di dati di validazione è un insieme separato di dati utilizzato per valutare le prestazioni del modello su dati non visti durante l'addestramento;
- val_accuracy: Rappresenta l'accuratezza del modello sul set di dati di validazione. In questo caso, l'accuratezza è del 70,82%.

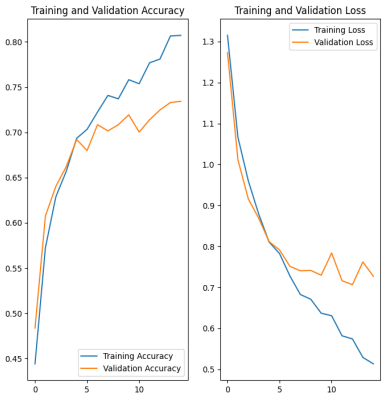
È interessante analizzare il valore dell'accuratezza e della perdita a fine addestramento. Il grafico che segue mostra un tipico esempio di accuratezza non ottimale, nel quale il modello ha raggiunto una percentuale solo del 60% circa sul set di convalida.



Si noti come l'accuratezza dell'addestramento aumenti linearmente nel tempo, mentre l'accuratezza della convalida si ferma a circa il 60%. Inoltre, è evidente la differenza di accuratezza tra addestramento e convalida rappresentata da un segnale di sovraccarico, il cosiddetto *overfitting*: questo fenomeno si verifica quando nell'addestramento si utilizza un numero limitato di campioni e, di conseguenza, il modello apprende dalle immagini rumori o dettagli indesiderati in misura tale, da influire negativamente su sue ulteriori prestazioni su campioni diversi.
Per i tre modelli sviluppati nella sperimentazione l'andamento dell'accuratezza è stato pressoché simile. Il grafico seguente rappresenta l'addestramento per il tipo *manoscritto*:



Questa invece l'accuratezza per i modelli dei *libri antichi e moderni*:



2. RISULTATI E PROSPETTIVE

I risultati della sperimentazione forniscono una panoramica incoraggiante delle prestazioni di ciascun modello sviluppato. È stata condotta una serie di test su un set di dati diversificato e implementata una procedura di validazione su un vasto set di dati.
In termini di accuratezza, il modello ha dimostrato una performance notevole, raggiungendo per le tre tipologie e per tutto il set di dati di convalida una media del 78% di corretta classificazione delle parti che compongono le strutture dei materiali

considerati. La curva di apprendimento ha mostrato una convergenza del modello, segno che con un ulteriore *tuning* degli iperparametri che governano il processo di training e la topologia di un modello machine learning sarà possibile raggiungere un valore di accuratezza più alto e, nel contempo, diminuire la perdita per quelle classi che hanno mostrato segni di incertezza. Si mostrano di seguito alcuni esempi degli output di addestramento (vd. Figg. 2, 3, 4, 5).

Figura 2. Visualizzazione di riconoscimento con errore

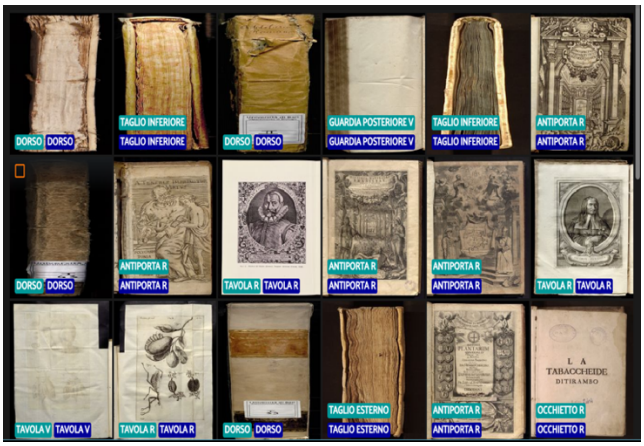


Figura 3. Esempio di Unique CORRECT predictions

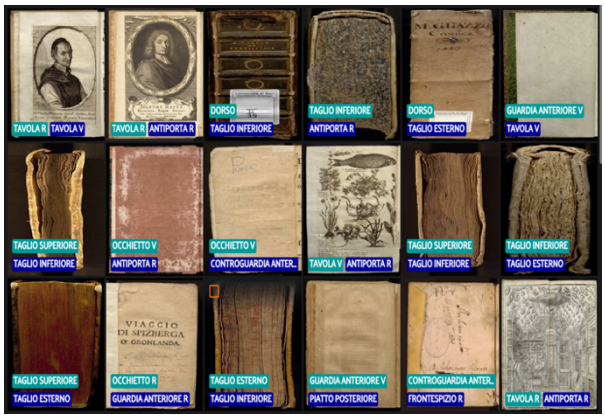


Figura 4. Esempio di INCORRECT predictions

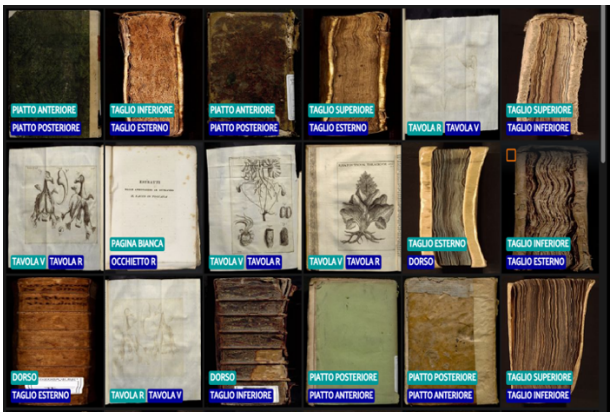


Figura 5. Annotazione errata data come training

Dataset	Dimensioni	Classi	Score medio
Antico	3k campioni	25	85%
Moderno	2k campioni	7	82%
Manoscritto	2k campioni	19	75%

Si rileva una distribuzione equa delle performance tra le diverse classi, che indica una buona capacità dei modelli di agire validamente su una varietà di categorie. Complessivamente, si conferma l'efficacia dell'approccio descritto, sottolineando il suo potenziale in contesti applicativi reali. Le classi da migliorare sono:

- Manoscritto: Piatto; Carta; Controguardia; Frontespizio;
- Antico: Piatto; Controguardia; Paratesto;
- Moderno: Pagina; Tavola.

Attualmente, la ricerca si sta concentrando sulle possibili soluzioni utili a portare al 100% l'accuratezza della classificazione in ognuno dei tre modelli sviluppati.

BIBLIOGRAFIA[10]

[1] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. '"ImageNet Classification with Deep Convolutional Neural Networks". Advances in Neural Information Processing Systems'. In Advances in Neural Information Processing Systems, edited by F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, Vol. 25, 2012. https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[3] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrej Zhmoginov, and Liang-Chieh Chen. 'MobileNetV2: Inverted Residuals and Linear Bottlenecks'. In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510–4520, 2018. https://openaccess.thecvf.com/content_cvpr_2018/papers/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.pdf

[4] Shanmugamani, Rajalingappaa. Deep Learning for Computer Vision: Expert Techniques to Train Advanced Neural Networks Using TensorFlow and Keras. Packt Publishing, 2018.

[10] Ulteriore ampia bibliografia scientifica relativa al tema discusso sarà curata nella redazione più ampia del presente contributo, di prossima pubblicazione. Nel frattempo, si rimanda ai seguenti link:
Academic literature on the topic 'Digital image classification'.
https://www.grafiati.com/en/literature-selections/digital-image-classification/
Academic literature on the topic 'Image classification techniques'.
https://www.grafiati.com/en/literature-selections/image-classification-techniques/

Uncovering the spread of lexical innovation in Italian tweets

Greta Franzini[1], Paolo Brasolin[2], Stefania Spina[3]
[1] Eurac Research, Institute for Applied Linguistics, Italy - greta.franzini@eurac.edu
[2] Eurac Research, Institute for Applied Linguistics, Italy - paolo.brasolin@gmail.com
[3] Università per Stranieri di Perugia, Italy - stefania.spina@unistrapg.it

ABSTRACT
This paper outlines ongoing research into lexical innovation in contemporary Italian in the context of social media. To date, the study has used a dataset of 5.32M timestamped and geotagged tweets extracted from the 2022 Italian timeline, yielding 720 emerging word forms. Here, we describe the reproducible pipeline developed to extract candidate neologisms from our dataset and introduce a custom tool to visualise the emergence and spread of candidate neologisms in the time-period under investigation.

KEYWORDS
Twitter; Italian; social media; lexical innovation; language change.

1. INTRODUCTION
The evolution of language is often propelled by lexical innovation [6, 13]. This phenomenon involves the generation of new words[1] and their assimilation into established lexical frameworks [10]. The process typically unfolds through distinct stages, starting with limited usage in specific contexts, progressing to wider adoption, and, in certain instances, formal inclusion in dictionaries [7, 11].
The creation of novel words is influenced by diverse linguistic processes such as derivation, compounding, transcategorisation, blending, semantic shifts, and borrowing from other languages. Social media platforms provide an avenue to study the emergence of new lexical forms in everyday discourse, offering real-time insights from a varied demographic, complete with geotagged text to examine geographical variations [9].
Our research into the language change mechanisms of contemporary Italian has so far examined Twitter interactions from the 2022 Italian timeline. First results have yielded a list of 720 emerging candidates classified into 14 categories of lexical creation. Here, we describe the reproducible pipeline we developed to identify new word forms in tweets and introduce a bespoke tool to dynamically visualise the emergence and spread of candidate neologisms in our dataset.

2. RELATED WORK
A key focus of previous research on lexical innovation in contemporary Italian [1, 3, 4, 15] has been the categorisation of the processes that lead to lexical creation. Traditionally, these include borrowing from other languages, forming new words from existing ones, changing grammatical categories and semantic shifting [21]. The *Osservatorio neologico della lingua italiana* (ONLI) [2] has meticulously monitored the appearance of new Italian words in newspapers, recording, to date, 2,986 new forms.
In recent years, social media has gained traction as a means of tracking new words in informal contexts [16, 19, 20]. By offering an unprecedented volume of conversational data from diverse speakers, social media facilitates robust evaluations of lexical creativity while enhancing the exploration of language variation and change [14, 17]. This approach has prompted studies to delve into the initial, less-documented phases of lexical innovation occurring after word creation [8, 9, 12] but before institutionalisation in dictionaries.

3. METHODOLOGY
Our research seeks to answer two research questions, namely *are Twitter conversations a reliable source to trace lexical innovation? And what are the linguistic processes leading to the creation of emerging words in Italian Twitter?* To this end, our work began with the sampling of 5.32M timestamped and geotagged tweets from the Italian timeline of 2022 using Twitter's advanced search query language [5]. These tweets were posted by 153,000 unique users for a total 71.5M tokens or 564M characters.

[1] The terms "word" and "form" are used interchangeably.

Tweets are complex structures. They may include geolocation information in the form of a latitude/longitude pair or a place. The latter is defined as an administrative region or a point of interest, identified by an ID, a country code, a geographical bounding box and additional metadata. In our dataset, 99.43% of tweets are associated with a place, 0.04% come with a latitude/longitude pair only, and 0.53% bear no geolocation information whatsoever[2]. Consequently, we chose to focus on tweets associated with a place as they cover almost the entire dataset. Of these, 91.77% contain places associated with an IT country code: we identified 34.8K unique places and computed the centroid of their bounding box, matching it with governmental data[3] to generate choropleth maps for geographical analysis. The remaining data includes 8.16% tweets related to places with other country codes and 0.07% tweets representing the entirety of Italy. As well as geolocation information, tweets contain an ID, a user ID, a timestamp, the full text, a list of "entities" and other metadata. An "entity" is a character range in the full text labelled with a type (*url*, *user mention*, *hashtag*, *symbol* or *media*) and other metadata. Recognising the value of entity metadata for the downstream tokenisation process, we incorporated them into the full text as delimiter markers by selecting a unique pair of Unicode code points for each entity type from the Private Use Area in the Basic Multilingual Plane.

We tokenised our corpus with the spaCy v3.6.1 Italian tokeniser. Tweets present a number of challenges for a standard tokeniser due to the widespread use of Unicode, variable casing, white-spacing and punctuation marks. To address these issues, we took several steps, including replacing emojis with spaces, converting the text to lowercase, trimming sequences of white-space and extending the tokeniser's infix matcher to identify sequences of commonly abused punctuation marks. Finally, our aforementioned entity annotation allowed us to wrap delimited regions in the text with spaces to help the tokeniser find their beginning; additionally, we defined a custom token matcher to catch any sequence bounded by our delimiter character pairs. Thanks to these adjustments, the tokeniser produced a minimal number of spurious tokens. Next, we filtered the output, eliminating tokens consisting of pure spaces, punctuation, numbers, and broken or non-existent handles (i.e., tokens beginning with @ but not marked as entities). We retained only hashtag entities. The process resulted in the extraction of 71.5M tokens (926K types).

To identify candidate neologisms, we employed two different strategies. The first is that adopted by [8, 9], which involves computing a measure of the degree to which the use of a token increases monotonically over time, and excluding tokens falling below a given positive threshold. The chosen measure, denoted as ρO, is the Spearman rank correlation coefficient between the daily occurrences of a token (normalised by the daily total token count) and the day number. We also computed the same measure on the daily users count ρU and allowed negative values because we noticed plausible patterns of new words use among them (e.g., an early peak followed by a lower plateau). Selecting forms matching any condition among $\rho O > 0.2$, $\rho O < -0.2$, $\rho U > 0.2$ or $\rho U < -0.2$, we obtained 6,737 candidate neologisms, which we defined as subset $\boldsymbol{A}$. In measuring monotonicity, this method might however still exclude other plausible patterns of new word use. We therefore employed an alternative, complementary strategy aimed at excluding usage patterns that we *do not* expect from emerging forms, such as accidental and sporadic phenomena (e.g., typos, inside jokes, etc.), excluded by setting lower bounds to the count of occurrences ($O > 9$) and unique users ($U > 9$); forms in use from the past or ending prematurely, excluded by setting lower bounds to the days of first ($A > 7$) and last ($Z > 351$) occurrence; and short-lived forms, excluded by establishing a lower bound to the usage lapse ($Z - A > 28$). The alternative subset, defined as $\boldsymbol{B}$, yielded 21,132 candidate neologisms (979 overlapping with $\boldsymbol{A}$). Combined, subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ count 26,890 candidate neologisms (2.90% of the total extracted forms). From these, we automatically removed 15,366 using a lexicon of 514K Italian forms [18] and were thus left with 11,524 candidates between hashtags (3,391) and non-hashtags (8,133) for manual annotation.

The manual annotation process was conducted by two authors of the present paper. We used AntConc to look up the forms' context and three online dictionaries for their attestation[4], categorising candidates as either "innovative" or "non-innovative" and resolving sporadic disagreements through negotiation. Next, we grouped innovative forms into one or more categories of lexical creation based on an adjusted ONLI typology scheme.

Finally, we produced choropleth maps and developed a custom tool to visualise the distribution of our innovative forms across Italy.

4. RESULTS AND DISCUSSION

Methodologically, our approach avoids Grieve's inherent bias towards monotonic patterns of emergence (by, for instance, accounting for forms that emerge early in the dataset but settle on a plateau lower than the initial peak), is conceptually

[2] This is possible because Twitter data can be redacted.

[3] https://www.istat.it/it/archivio/222527; https://github.com/openpolis/geojson-italy

[4] https://www.garzantilinguistica.it/; https://slengo.it/; https://www.treccani.it/vocabolario/

simpler and computationally more efficient having benchmarked it to be 50x faster.

The annotation of non-hashtag and hashtags forms differed with respect to innovation criteria and ONLI classification. Starting with non-hashtags forms, we excluded attestations, typographical errors caused by key proximity, popular terms (e.g., *bimbominchia*), established loanwords (e.g., *foliage*, *sponsorship*), adapted loanwords (e.g., *followo*, *crashare*), infrequently used foreign words (e.g., *smoothie*, *veggie*), acronyms (e.g., *PTSD*), regionalisms and regional variants (e.g., *annassero*, *ciolla*), gender-inclusive graphic variants (e.g., *cittadinə*), and nicknames (e.g., *pupone* for footballer Francesco Totti). The annotation resulted in 347 emerging forms, grouped into 14 categories of lexical creation. The most prolific categories are: orthographic variation used for emphasis (e.g., *pikkolo*), fun and sarcasm (e.g., *scienzah*), to shorten existing words or hide online conversation (e.g., *f4scist4*); univerbation (e.g., *stemmerde*); suffixation, used for intensification (e.g., *adorissimo*), pejorative connotation (e.g., *cinesata*) or augmentation (e.g., *soggettone*); loanwords, mostly borrowed from English (e.g., reminder); and portmanteau forms or blends, typically serving to amuse (e.g., *assurdistan*, *lettamaio*). Of these 347 forms, 22 are now part of the Zingarelli 2024 Italian monolingual dictionary, a 2023 publication superseding the previous edition with 250 new words and 750 new multi-word forms[5]. These 22 forms are obtained through suffixation, adapted and direct borrowing, prefixation, deonymic derivation, transcategorisation, and creation of portmanteau forms. Interestingly, forms resulting from orthographic variation, a common source of lexical innovation on social media, do not seem to make it into the dictionary, suggesting that spelling changes are not a strong criterion for lexicographic acceptance. Grammatically, most of our (now) institutionalised forms are nouns (12), with a few serving as both nouns and adjectives (4), followed by verbs (3) and adjectives (3). This suggests that, in the context of Twitter, nouns derived through suffixation are the most likely candidates for inclusion in dictionaries. As for hashtag annotation, to address biases introduced by forced univerbation and English dominance in detecting patterns of lexical creation, our analysis follows both objective and subjective criteria. We narrowed our selection by filtering out hashtags used by nine or fewer users, hashtags containing proper names (including but not limited to people, places, organisations, brands, sports teams, events, festivities, videogames, music bands, concerts, films, TV shows and political stances/movements); (combinations of) years, days of the week, times, and numbers; short-lived hashtags relating to a specific incident or time interval; univerbated hashtags with little to no probability of inclusion in lexical resources. The remaining hashtags were then categorised into single and univerbated forms. While the annotation of the former mirrored that of non-hashtag forms, in the annotation of univerbated hashtags (e.g., *#andratuttobene*, *#booklover*) we only considered those that intuitively seem likely to establish themselves as new non-hashtag forms in Italian social media communication or be acknowledged in authoritative lexical resources (e.g., *#avantitutta*, *#oldschool*). Of the emerging hashtags identified, 75% are loanwords.

We visualise forms belonging to subset ***A*** as choropleth maps and forms subset ***A*** and ***B*** using a custom tool. The choropleth maps, shown in Figure 1, display the *regional* word instances per million tokens of forms from four different ONLI categories. The map of *gomblotto* shows widespread orthographic variation in almost all regions, particularly in Lombardy. Univerbated forms are thinly spread, with regional peaks like *miraccomando*, which is particularly popular in Lombardy. Loanwords like *flexo* are prominent in the west but absent in the lower east, while *fattoni* is used in northern regions but is absent from the southern-east part of the country and the islands. Our custom tool, shown in Figure 2, generates a more accurate and interactive alternative visualisation from latitude/longitude pairs and timestamps for a geo-temporal analysis of the emerging forms. The tool currently takes a four-column CSV data file as input (*timestamp*, *latitude*, *longitude*, *word*) and allows users to sort one or multiple forms by time-period and/or geographical area.

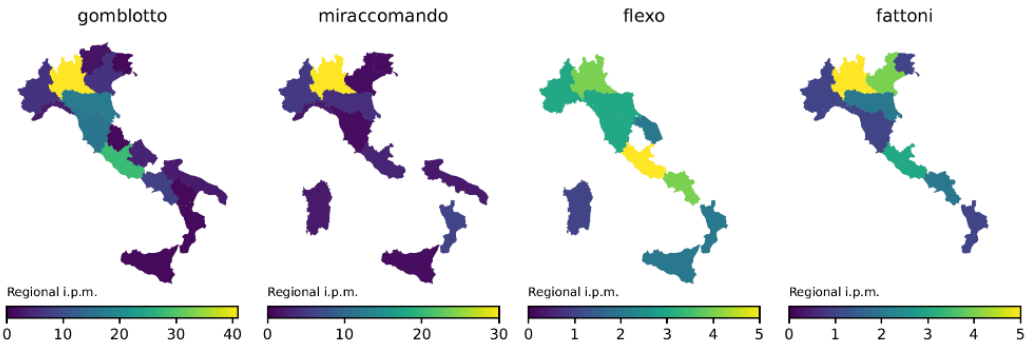


*Figure 1. Choropleth maps of Italy showing regional instances per million tokens of gomblotto, miraccomando, flexo and fattoni (forms taken from subset **A**)*

[5] https://www.zanichelli.it/ricerca/prodotti/lo-zingarelli-2024. This dictionary is generally regarded as the most up-to-date neologism lexicon for Italian.

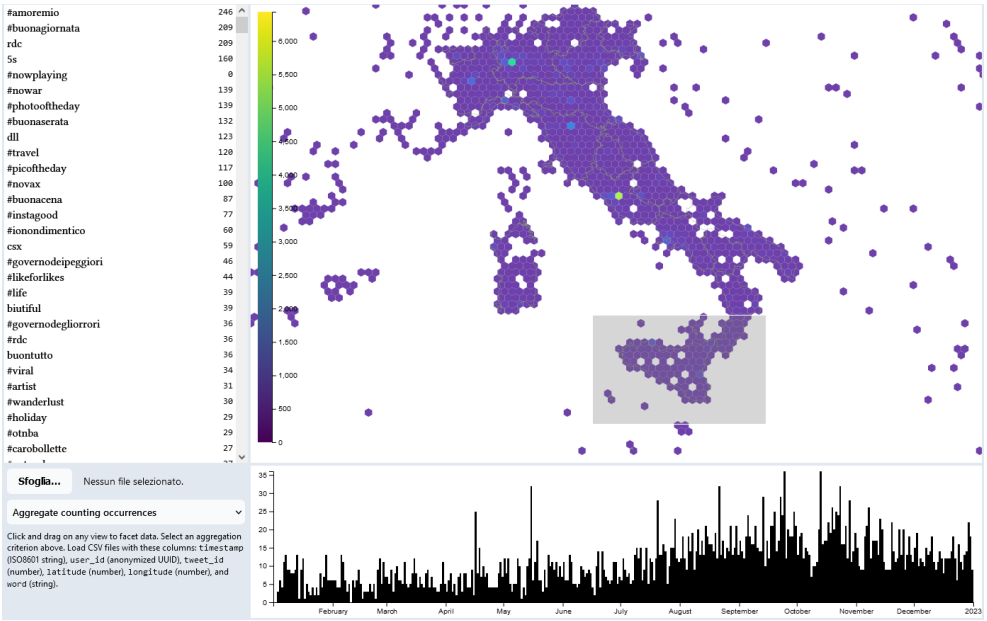


Figure 2. Screenshot of Frustum*, our custom tool, developed to interactively visualise the geo-temporal behaviour of emerging forms in our dataset (forms taken from subset **A** and **B** combined)*

5. CONCLUSION

This study investigates lexical innovation in Italian Twitter, revealing that the emergence of new words seems to be driven more by creativity, entertainment and a sense of belonging to a community rather than a need for novel terms to describe new concepts, situations or events. The 720 identified forms primarily serve functions related to irony, intensification and emphasis, and some coined expressions may extend beyond spoken discourse to online communication and media use. Overall, this study shows that Twitter and social media in general can be a reliable source for the study of lexical innovation, as they provide large amounts of data produced by ordinary speakers in their written and informal interactions. These data can effectively complement those traditionally used in the study of lexical innovation, such as data taken from the press. Furthermore, the study shows that, alongside traditional mechanisms of lexical innovation, Twitter interactions reveal a widespread use of spelling variation as a means of coinage of emerging forms, for it is capable of conveying different nuances of meaning.

While our one-year timeframe captures rapid linguistic phenomena on Twitter, it likely misses slower-spreading forms and, with them, other patterns of lexical creation. Follow-up work might, therefore, extend the analysis to additional timelines (and alternative micro-blogging platforms, given Twitter's recent developments) focussing on the proliferation dynamics and potential institutionalisation of these candidate neologisms.

Our study contributes to the state of the art in the field with the largest study on Italian yet, an annotated dataset, a computationally-lighter and reproducible pipeline to automatically process micro-blogging posts for candidate neologisms, and a user-friendly visualisation tool to interactively browse geolocated data. The annotated datasets, the maps and the code resulting from this study are all available at https://github.com/breviloquia-italica.

REFERENCES

[1] Adamo, Giovanni, and Valeria Della Valle, eds. 'Neologismi quotidiani. Un dizionario a cavallo del Millennio'. In *Lessico intellettuale europeo*, Vol. 95. Olschki, 2003.

[2] Adamo, Giovanni, and Valeria Dalle Valle. 'Osservatorio neologico della lingua italiana. Lessico e parole Nuove 'italiano'. In *1*. Roma: ILIESI-CNR, 2019. http://www.iliesi.cnr.it/scheda.php?id=253&cl=I/TS.

[3] Adamo, Giovanni, and Valeria Della Valle, eds. 'Innovazione lessicale e terminologie specialistiche', *Lessico Intellettuale Europeo*, Vol. 92. Olschki, 2003.

[4] Adamo, Giovanni, and Valeria Della Valle. 'Neologismi (Parole nuove dai giornali 2008-2018)'. In *Istituto dell'Enciclopedia Treccani*. Roma, 2018.

[5] Brigadir, Igor. 'Advanced Search On Twitter', 2023. https://github.com/igorbrigadir/twitter-advanced-search

[6] Croft, William. Explaining Language Change: An Evolutionary Approach. Harlow: Pearson Education, 2000.

[7] Fischer, Roswitha. Lexical Change in Present-Day English: *A Corpus-Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Vol. 17. Language in Performance. Tübingen: G. Narr, 1998.

[8] Grieve, Jack, Nini Andrea, and Diansheng Guo. 'Analyzing Lexical Emergence in Modern American English Online'. *English Language & Linguistics* 21, no. 1 (2016): 99–127.

[9] Grieve, Jack, Andrea Nini, and Diansheng Guo. 'Mapping Lexical Innovation on American Social Media'. *Journal of English Linguistics* 46, no. 4 (1 December 2018): 293–319. https://doi.org/10.1177/0075424218793191

[10] Ježek, Elisabetta. *The Lexicon: An Introduction*. Oxford Textbooks in Linguistics. Oxford, New York: Oxford University Press, 2016.

[11] Kerremans, Daphné. A Web of New Word. Frankfurt am Mai: Peter Lang, 2015.

[12] Kershaw, Daniel, Matthew Rowe, and Patrick Stacey. 'Towards Modelling Language Innovation Acceptance in Online Social Networks.' *In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 553–562. WSDM '16. New York, NY, USA: Association for Computing Machinery, 2016. https://doi.org/10.1145/2835776.2835784

[13] Labov, William. *Principles of Linguistic Change*, Volume 2: Social Factors. Vol. 2–3. Oxford: Wiley-Blackwell, 2001.

[14] Laitinen, Mikko, Masoud Fatemi, and Jonas Lundberg. 'Size Matters: Digital Social Networks and Language Change'. *Frontiers in Artificial Intelligence*, 46, 3 (20 July 2020). https://doi.org/10.3389/frai.2020.00046

[15] Marri, Fabio. 'The Neologisms inside and Outside the Recent Repertoires'. *Quaderns d'Italià* 23, no. 11 (5 December 2018). https://revistes.uab.cat/quadernsitalia/article/view/v23-marri

[16] Rodríguez Arrizabalaga, Beatriz. 'Social Networks: A Source of Lexical Innovation and Creativity in Contemporary Spanish'. *Languages*, 138, 6, no. 3 (September 2021): 1–22. https://doi.org/10.3390/languages6030138

[17] Spina, Stefania. *Fiumi di parole. Discorso e grammatica delle conversazioni scritte in Twitter.* Aracne, 2019.

[18] Spina, Stefania. 'Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione'. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-It 2014*, edited by Roberto Basili, Lenci Alessandro, and Bernardo Magnini, 1:354–359. Pisa: Pisa University Press, 2014.

[19] Tarrade, Louise, Jean-Philippe Magué, and Jean-Pierre Chevrot. 'Detecting and Categorising Lexical Innovations in a Corpus of Tweets.' *Psychology of Language and Communication* 26, no. 1 (1 January 2022): 319–329. https://doi.org/10.2478/plc-2022-15

[20] Würschinger, Quirin. 'Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter.' *Frontiers in Artificial Intelligence*, 648583, 4 (1 November 2021). https://doi.org/10.3389/frai.2021.648583

[21] Zolli, Paolo. *Come nascono le parole italiane*. Milano: Rizzoli, 1989.

ORGANIZZAZIONE DELLA CONOSCENZA CON SEMANTIC WEB

Affinare il contesto: estrazione di informazioni strutturate per l'arricchimento dei contesti archivistici

Lucia Giagnolini[1], Paolo Bonora[2], Francesca Tomasi[3]
[1] Università di Bologna, Italia - lucia.giagnolini2@unibo.it
[2] Università di Bologna, Italia - paolo.bonora@unibo.it
[3] Università di Bologna, Italia - francesca.tomasi@unibo.it

ABSTRACT[1]

Gli strumenti di corredo archivistici in LOD sono spesso solo parzialmente capaci di esprimere il vero potenziale informativo dei dati, a causa della molteplicità di campi non strutturati presenti nelle descrizioni dei complessi documentari. La presenza di numerose sezioni *literal*, ovvero a testo pieno, limita da un lato la possibilità di interrogazioni a base semantica e dall'altro non consente l'apertura ai numerosi contesti latenti che tali porzioni di testo non strutturato veicolano. Si intende allora qui presentare una metodologia per acquisire nuova conoscenza dai dati, aprendoli al dialogo con nuovi contesti impliciti. A questo scopo, si è valutato quanto possano essere utili alcuni tool ad oggi disponibili per acquisire e restituire informazione strutturata rispetto alle descrizioni archivistiche. Attraverso un caso di studio del Sistema Archivistico Nazionale in LOD si sono analizzate le potenzialità di TINT, FRED e di ChatGPT nell'estrarre informazione morfosintattica, lessicale o semantica dai dati archivistici, riflettendo al contempo sulla possibilità di far dialogare il grafo di conoscenza nativo e il grafo risultante dall'analisi, e documentando gli atti interpretativi emersi.

PAROLE CHIAVE

Linked Open Data; archivi; information retrieval; supervised annotation; contesti.

1. PREMESSA

Un archivio è composto da elementi che, ciascuno con le proprie caratteristiche, permettono di fornire una rappresentazione stratificata del complesso di attività ed entità coinvolte nella produzione dei documenti. Per renderlo manifesto, è necessaria una chiara esposizione della rete di relazioni che collega le singole parti tra loro e il fondo stesso con i molteplici contesti di riferimento [5]. Se, da un lato, le descrizioni basate sullo standard metodologico ISAD(G) hanno permesso una funzionale formalizzazione e strutturazione dell'atto descrittivo, dall'altro è ormai largamente appurato che l'applicazione dello standard ha determinato rappresentazioni primariamente impostate su relazioni gerarchiche – dunque strettamente verticali e scarsamente permeabili ai contesti – valorizzando solo marginalmente la restituzione in termini di legami orizzontali [5, 17]. Anche per questo motivo, da ormai più di un decennio, le istituzioni dell'ambito GLAM si sono avvicinate al paradigma dei Linked Open Data (LOD), che "ha richiesto di rivedere sistematicamente le informazioni riportate nelle descrizioni archivistiche e nelle schede catalografiche (destrutturando e ristrutturando tipologia, granularità e precisione), così da superare gli schemi documento-centrici della descrizione con approcci data-centrici, che valorizzano le relazioni con il contesto" [8]. Nella migrazione di inventari archivistici tradizionali in LOD, blocchi informativi estremamente rilevanti – come la nota biografica, la storia archivistica e i criteri di riordinamento – vengono spesso trasposti esclusivamente come corpose stringhe di caratteri *literal*, ossia in testo pieno. Si tratta di campi testuali estremamente ricchi di informazioni, che potrebbero essere strutturate in modo più organizzato e funzionale, rappresentando "il carburante indispensabile a far decollare il razzo dell'integrazione multicontestuale" [15: 5]. Infatti, il Semantic Web non ha cambiato l'approccio tradizionale delle istituzioni alla descrizione, ma ha enfatizzato la necessità di adottare una semantica esplicita, in modo tale da consentire una interoperabilità basata sull'impiego di modelli concettuali, facilitando il riuso dei dati [14]. Ogni nuova asserzione espressa in forma di tripla diventa generatrice di inferenza e di nuova informazione: più i contesti di appartenenza di queste asserzioni crescono e si intersecano, più la rete semantica si arricchisce e diventa informazione classificata [10]. In altre parole, i contenuti testuali di campi descrittivi, espressi come sequenze di stringhe e rappresentati in forma aggregata attraverso semplici nodi di tipo *literal* – pur mantenendo la loro unitarietà nel campo descrittivo – potrebbero essere esplicitati attraverso nuove triple. Ogni nuova tripla diventerebbe portatrice di una componente informativa specifica presente nel testo aggregato come, ad esempio, le attestazioni di istituzioni, persone, eventi e coordinate spazio-temporali. L'estrapolazione della specifica semantica del dato preesistente, attraverso la creazione di triple attestanti relazioni più o meno esplicite nel testo, consentirebbe un arricchimento significativo della *knowledge base*

[1] L. Giagnolini ha curato le sezioni 1 e 2, con P. Bonora la sezione 3; le conclusioni sono il risultato di una riflessione collettiva degli autori.

contestuale, permettendo, fra l'altro, un notevole miglioramento delle operazioni di ricerca e, potenzialmente, anche la disambiguazione delle entità citate.
Le operazioni di estrazione delle entità presenti nel testo e l'attribuzione della semantica di relazione tra di esse rappresentano a tutti gli effetti un atto interpretativo del contenuto testuale [6]. È necessario, quindi, che le nuove triple, indipendentemente dal fatto che siano il risultato di un processo di estrazione supervisionato o meno, vengano esplicitamente individuate come il risultato di una nuova attività di analisi, distinta dall'azione di descrizione archivistica, che ha prodotto il record originario. Le triple finalizzate all'arricchimento del dato, dovranno quindi essere corredate da una serie di ulteriori triple che ne dichiarino espressamente l'origine, le modalità di produzione e, in ultima istanza, l'attribuzione di responsabilità. Ovvero, ne esplicitino la cosiddetta *provenance* [14].
In sintesi, l'adozione del paradigma Linked Open Data (LOD) ha aperto la strada a una nuova prospettiva nella descrizione archivistica [12], ma per superare davvero le limitazioni gerarchiche e favorire la creazione di una knowledge base semanticamente ricca occorre sfruttare al meglio i campi di testo e strutturarne i contesti latenti, aggiungendo adeguata documentazione al processo di produzione di nuova conoscenza.

2. METODOLOGIA E WORKFLOW

Il tentativo di "stabilire se e in che misura le tecniche e le tecnologie di gestione del testo possano potenziare i nostri strumenti nel rispetto del contesto, arricchendoli di appigli informativi" [15: 7], si traduce nel comprendere come impiegare gli strumenti ad oggi disponibili per acquisire informazione strutturata dalle descrizioni archivistiche. A questo scopo, è necessario chiarire gli step del processo di estrazione dell'informazione, ovvero elaborare un modello di workflow che sia capace di contemplare tanto l'esigenza di definire il tipo di analisi da delegare allo strumento, quanto la necessità di valutare degli esiti della sua applicazione. L'approccio che proponiamo per l'implementazione del processo è articolato nei seguenti punti:

1. Selezionare il tipo di atto interpretativo che si delega allo strumento (ad esempio, analisi morfosintattica, lessicale o semantica) a seconda dei contenuti da analizzare.
2. Individuare le tecnologie e le rispettive implementazioni in funzione del tipo di atto interpretativo atteso (ad esempio, da tecniche NLP elementari al deep learning e LLM).
3. Definire il modello di valutazione dell'esito e della qualità dell'atto interpretativo automatico, dove per qualità si intende "la possibilità di attingere a dati ragionevolmente affidabili, perché parte di un contesto che li giustifica e li spiega" [16: 10]. Gli output dell'atto interpretativo dovranno, infatti, essere vagliati e selezionati da un esperto di dominio per essere ritenuti validi.
4. Individuare il modello di rappresentazione e sedimentazione della conoscenza estratta nell'ottica di una struttura semanticamente controllata e interoperabile dal punto di vista dell'acceso al dato (ad esempio, RDF in prospettiva LOD).
5. Modellare i criteri e le modalità di acquisizione della conoscenza estratta in funzione della capacità espressiva del relativo modello descrittivo (ad esempio, Dublin Core, RiC-O, SAN LOD) e dei criteri redazionali. A questo scopo andrà definito un modello di attestazione della *provenance* del dato che espliciti il tipo di atto interpretativo, lo strumento e il processo utilizzato per ottenerlo, le metriche di valutazione (ad esempio, *recall* e *precision*) e il riferimento al supervisore scientifico (ossia l'attribuzione di responsabilità).
6. Valutare le strategie per mettere in relazione il dato analizzato nel sistema nativo e la serie di triple esito dell'atto interpretativo.
7. Modellare l'interazione utente-sistema in termini di processo operativo e di interfacce, ovvero individuare strategie di information visualization che consentano al contenuto informativo estratto di essere adeguatamente presentato e gestito.
8. Valutare potenziali modalità di rinforzo dello strumento esterno per migliorarne le prestazioni (ad esempio, training set NLP, miglioramento del prompt per ChatGPT).

In questa formulazione, il processo è sufficientemente astratto da poter essere applicato a contesti diversi e obiettivi di estrazione dell'informazione operanti a molteplici livelli: dall'analisi della superficie lessicale all'interpretazione della semantica del testo.

3. PROOF OF CONCEPT

Per presentare possibili esiti dell'approccio metodologico così illustrato, è possibile effettuare sperimentazioni con strumenti immediatamente disponibili che, anche se non ancora integrati all'interno di un concreto workflow, non presentano barriere d'accesso tali da impedirne un utilizzo dimostrativo.

Ad esempio, possiamo analizzare il campo denominato "descrizione" di una scheda “soggetto produttore” del Sistema di Archiviazione Nazionale (SAN), corrispondente alle proprietà “dc:description” e “abstract” del tracciato schema SAN. Individuiamo nella scheda SAN dedicata alla nota biografica di Andrea Costa[2] (1851-1910) il testo campione, prendendo in analisi il primo paragrafo della descrizione:

> Nasce a Imola il 29 novembre 1851 da Pietro e Rosa Tozzi in una famiglia cattolica praticante e di modeste condizioni. Il giorno successivo è battezzato nella cattedrale di S. Cassiano con i nomi di Andrea, Antonio e Baldassarre e suo padrino è Orso Orsini. Frequenta le scuole elementari gestite da un sacerdote e negli anni scolastici 1866-1867 e 1867-1868 frequenta la scuola tecnica comunale con Gaetano Darchini, Luigi Sassi e Angelo Negri. Negli anni scolastici 1868-1869 e 1869-1870 frequenta il liceo come uditore per le lezioni di letteratura italiana e latina. Il 15 dicembre 1870 si iscrive alla facoltà di filosofia e belle lettere dell'Università di Bologna come "studente libero" non avendo la possibilità di pagare le regolari tasse di ammissione e per mantenersi si impiega come scrivano in un'agenzia di assicurazioni imolese. Lì un impiegato, Paolo Renzi, lo associa, o almeno lo avvicina, all'Internazionale. A Imola e a Bologna compie il suo noviziato, nell'atmosfera che presto si accenderà degli entusiasmi per la Comune, e nel contatto con Carducci, che lo predilige fra i suoi allievi.

Stante l’obiettivo di identificare, isolare ed estrarre informazioni relative al soggetto contenute nella nota biografica, abbiamo selezionato tre strumenti progettati per operare sui tre livelli di analisi del testo (passo n. 1 e n. 2 del workflow): annotazione morfosintattica e NER (Tint[3] [1]); estrazione del significato su base lessicale e dei relativi nessi sintattici (FRED[4] [9]); interpretazione del testo su base statistica (ChatGPT[5]). I tre strumenti sono stati selezionati in considerazione del grado di maturità, dell’immediatezza d’uso, delle possibilità di ulteriore affinamento dell’addestramento. La sperimentazione, lungi dall’essere definitiva, mira alla semplice verifica della percorribilità dell’approccio basandosi su soluzioni terze, prescindendo dallo sviluppo in proprio di componenti dedicati.

L’applicazione di Tint per l’analisi del primo paragrafo della nota biografica ha permesso di individuare, tramite NER, i nomi di organizzazioni, luoghi e persone (vd. Fig. 1), così come le dipendenze sintattiche del testo, classificando in modo automatico le parti del discorso[6] (vd. Fig. 2).

Figura 1. Entità riconosciute nel testo e relativa classificazione

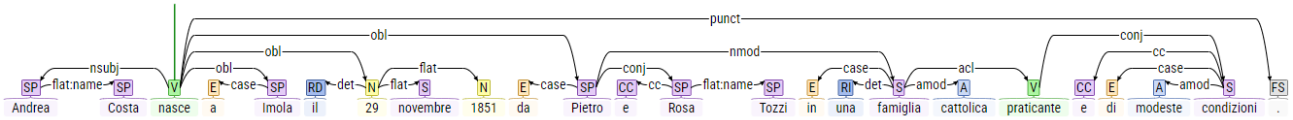


Figura 2. Grafo delle dipendenze sintattiche relative al primo periodo

[2] http://dati.san.beniculturali.it/SAN/produttore_IT-ER-IBC_san.cat.sogP.66756
[3] Per le finalità di questo intervento, è stata utilizzata la versione “Online demo” disponibile al link: https://dh.fbk.eu/tint-demo/
[4] Per le finalità di questo intervento, è stata utilizzata la versione “Online demo” disponibile al link: http://wit.istc.cnr.it/stlab-tools/fred/demo/
[5] Per le finalità di questo intervento è stata utilizzata la versione GPT-3.5 https://chat.openai.com/
[6] Per visualizzare il risultato integrale dell’analisi effettuata tramite Tint, v. Giagnolini, Lucia, and Paolo Bonora. “Affinare Il Contesto: Estrazione Di Informazioni Strutturate Per L’arricchimento Dei Contesti Archivistici. Risultati Dell'analisi Effettuata Con Tint”. figshare, January 31, 2024. https://doi.org/10.6084/m9.figshare.25119116.v4

L'applicazione di FRED ha prodotto una rappresentazione unificata e formalizzata in grafo di fatti e concetti espressi dal testo in linguaggio naturale[7] come, ad esempio, l'interpretazione delle condizioni di nascita di Andrea Costa (vd. Fig. 3).

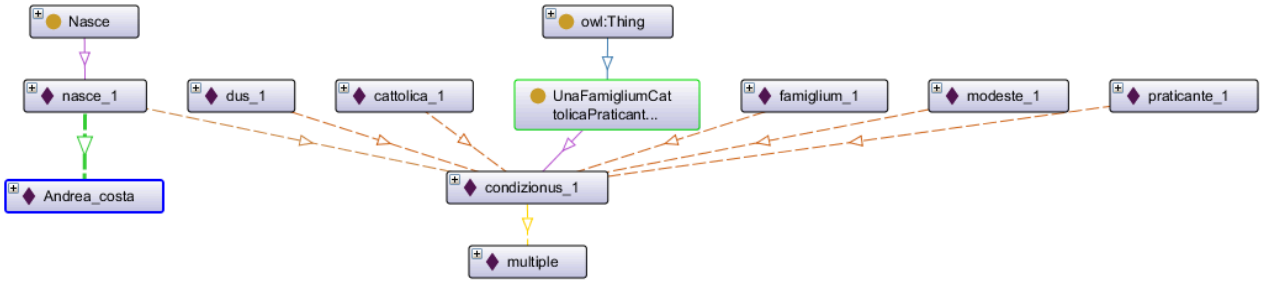


Figura 3. Grafo dell'interpretazione delle condizioni di nascita elaborata da FRED

ChatGPT 3.5 è stata utilizzata per sondare l'emersione di concetti più complessi, potenzialmente variabili a seconda dell'obiettivo e della strutturazione del prompt. A scopo esemplificativo, il prompt "*Elenca le entità che emergono da questo testo*" ha fornito i seguenti risultati:

- **Luogo di nascita**: Imola (29 novembre 1851)
- **Genitori**: Pietro e Rosa Tozzi
- **Ambiente familiare**: Famiglia cattolica praticante e di modeste condizioni
- **Atti religiosi**: Battezzato nella cattedrale di S. Cassiano con i nomi di Andrea, Antonio e Baldassarre; padrino Orso Orsini
- **Educazione**: Scuole elementari gestite da un sacerdote; frequenta la scuola tecnica comunale con Gaetano Darchini, Luigi Sassi e Angelo Negri; frequenta il liceo come uditore per le lezioni di letteratura italiana e latina
- **Università**: Iscritto alla facoltà di filosofia e belle lettere dell'Università di Bologna il 15 dicembre 1870 come "studente libero"
- **Situazione finanziaria**: Non ha la possibilità di pagare le regolari tasse di ammissione all'università; si mantiene lavorando come scrivano in un'agenzia di assicurazioni imolese
- **Influenza politica**: Associato o avvicinato all'Internazionale da Paolo Renzi, un impiegato nell'agenzia di assicurazioni
- **Contesto storico/politico**: Noviziato a Imola e a Bologna nell'atmosfera degli entusiasmi per la Comune; contatto con Carducci, che lo predilige fra i suoi allievi.

I tre strumenti producono output con gradi di finitura crescenti. Ai risultati prodotti sia con Tint che da FRED devono essere applicati algoritmi di estrazione in base a criteri semantici [3] per ottenere informazione strutturata da sottoporre alla valutazione degli esperti di dominio (passo n. 3 del workflow). Più immeditatamente strutturato è, invece, l'output prodotto dall'LLM (ChatGPT) che comunque dovrà essere allineato alla struttura del modello concettuale (passo n. 4).
Notiamo, inoltre, che l'estrazione delle informazioni di carattere temporale contenute nel testo, fondamentali per l'arricchimento dei contesti, prodotta dai tre strumenti oggetto della sperimentazione risulta lacunosa. Per superare questo limite, è possibile prevedere l'integrazione di strumenti dedicati come i Time Taggers [13] nel passo n.2 o un ulteriore raffinamento del prompt di ChatGPT[8].
A questo punto, occorre tenere presente che le informazioni estratte con questo approccio potrebbero non essere direttamente reintegrabili nella knowledge base d'origine, per limitazioni ontologiche della stessa (passo n. 6). Per quanto riguarda i dati strutturati estratti dalla nota biografica di Andrea Costa, ad esempio, nell'ambito del modello proposto dall'ontologia SAN LOD[9], emerge l'assenza di classi e proprietà in grado di rappresentare adeguatamente le informazioni estratte. Dunque, dal momento che la rappresentatività del modello da cui sono state acquisite le informazioni potrebbe diventare un ulteriore ostacolo all'esplicitazione dei contesti latenti, è più opportuno optare per un approccio che astragga dalle infrastrutture specifiche. Gli esiti dell'analisi possono confluire in un grafo in grado di rappresentare le nuove triple

[7] Per visualizzare il risultato integrale dell'analisi effettuata tramite FRED, v. Giagnolini, Lucia, and Paolo Bonora. "Affinare Il Contesto: Estrazione Di Informazioni Strutturate Per L'arricchimento Dei Contesti Archivistici. Risultati Dell'analisi Effettuata Con Fred". figshare, April 3, 2024. https://doi.org/10.6084/m9.figshare.25534225.v1
[8] https://platform.openai.com
[9] http://dati.san.beniculturali.it/lode/aggiornato.htm#d4e2193

nella loro massima granularità in una estensione stand-off. La nota biografica – o altri campi descrittivi dell'infrastruttura di riferimento – possono diventare l'oggetto di un asserto che stabilisce il legame tra la knowledge base d'origine e il grafo derivante dall'interpretazione del testo, sulla base, ad esempio, del modello proposto dalla Web Annotation Ontology[10] (vd. Fig. 4).

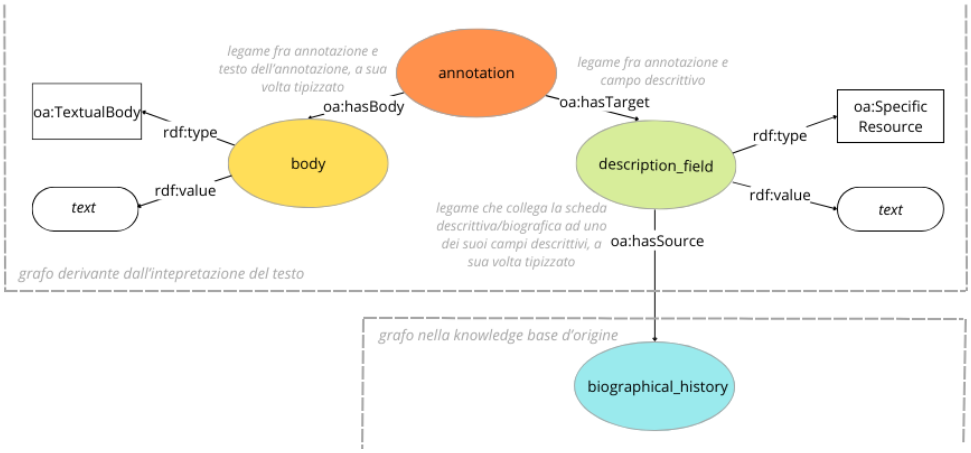


Figura 4. Rappresentazione del legame tra la knowledge base d'origine e il grafo derivante dall'interpretazione del testo

La Web Annotation Ontology permetterebbe di dar conto della *provenance* degli atti interpretativi e della loro potenziale molteplicità – sia in termini di strumenti che di esperti di dominio – con un conseguente ampliamento delle possibilità di esplicitazione dei contesti alla base dei nuovi grafi (passo n. 5) [6, 7]. Tuttavia, occorrerà estendere l'ontologia in modo tale da veicolare adeguatamente le informazioni estratte, a seconda delle esigenze di rappresentazione.
Sarà poi necessario individuare forme di visualizzazione dell'informazione così rappresentata, capaci di restituire all'utente nuova conoscenza attraverso un accesso sapiente alla molteplicità dei contesti che emergono dai processi di analisi (passo n. 7).

4. CONCLUSIONI E PROSPETTIVE

Questo contributo intende mettere in luce il ruolo fondamentale della destrutturazione del dato testuale per consentire un recupero semantico efficace dell'enorme e preziosissima mole di inventari archivistici pubblicati in rete. Infatti, il suo obiettivo non è valutare le performance dei singoli strumenti, bensì presentare una proposta di approccio metodologico per l'estrazione automatica di informazioni strutturate dalle descrizioni archivistiche. Ciò non toglie che, in una prospettiva di sviluppo, non ci si potrà esimere da un'analisi comparata approfondita dell'efficacia di questo approccio in termini di quantità e qualità dell'informazione estratta. In questi termini, le prime sperimentazioni, precedentemente illustrate, sembrano fornire risultati benauguranti, anche in considerazione del fatto che l'automazione della procedura non esaurisce la capacità interpretativa dell'essere umano, ma assume una funzione coadiuvante e documentale della stessa.
Le prospettive di lavoro si aprono soprattutto verso lo sviluppo e l'applicazione dei LLM [4, 11]: occorrerà determinare quali e quante operazioni di analisi siano effettivamente in grado di effettuare e con che grado di raffinatezza, anche per stabilire se i risultati forniti da tecniche NLP siano già superati o superabili, ragionando su come affinare gli strumenti utilizzati o il loro addestramento (passo n. 8). Il passo successivo si tratteggia sulla modellazione ed estrazione di concetti che vadano "oltre l'identificazione delle entità, enfatizzando sistemi di relazioni sempre più larghi e suggerendo di affiancare alla consolidata multilivellarità una multidimensionalità capace di rendere 'visibili' le idee o i fatti di cui le diverse entità sono veicoli" [16: 5]. Notiamo, infatti, che i risultati ottenuti attraverso l'utilizzo di ChatGPT offrono una gamma di informazioni che superano l'individuazione delle entità canoniche. Tuttavia, stabilizzarne e strutturarne i dati risultanti risulta notevolmente più complesso rispetto all'adozione di tecniche più tradizionali come la NER. Questo sottolinea l'importanza di ulteriori ricerche e sviluppi nel campo sia della *knowledge graph generation* dal linguaggio naturale che dell'uso delle classificazioni prodotte dagli LLM, al fine di aumentare il potenziale di tali modelli per la massimizzazione del contenuto informativo delle descrizioni archivistiche [2, 11].

[10] https://www.w3.org/ns/oa

5. RINGRAZIAMENTI

Contributo parzialmente finanziato dall'Unione europea - Next Generation EU, investimento I.4.1 Borse PNRR Patrimonio Culturale, Decreto Ministeriale n. 351 del 9 aprile 2022.

BIBLIOGRAFIA

[1] Aprosio, Alessio Palmero, e Giovanni Moretti. «Tint 2.0: An All-Inclusive Suite for NLP in Italian». In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-It. 10-12 dicembre 2018*, a cura di Elena Cabrio, Alessandro Mazzei, e Fabio Tamburini, 311–17. Torino: Accademia University Press, 2018. https://doi.org/10.4000/books.aaccademia.3571

[2] Babaei Giglou, Hamed, Jennifer D'Souza, e Sören Auer. «LLMs4OL: Large Language Models for Ontology Learning». In *The Semantic Web – ISWC 2023*, a cura di Terry R. Payne et al., 408–427. Cham: Springer Nature Switzerland, 3023. https://doi.org/10.1007/978-3-031-47240-4_22

[3] Bonora, Paolo, e Angelo Pompilio. «Automatic Extraction of Opera Character Characteristics through Lexical-Syntactic Patterns». *Umanistica Digitale* 5, fasc. 10 (gennaio 2021): 193–210. https://doi.org/10.6092/issn.2532-8816/12426

[4] Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, e Julia Noordegraaf. «Archives and AI: An Overview of Current Debates and Future Perspectives». *Journal on Computing and Cultural Heritage* 15, fasc. 1 (14 dicembre 2021): 4:1-4:15. https://doi.org/10.1145/3479010

[5] Damiani, Concetta. «Archival Description and Conceptual Transversality». *JLIS.It* 13, fasc. 3 (15 settembre 2022): 154–161. https://doi.org/10.36253/jlis.it-485

[6] Daquino, Marilena, e Francesca Tomasi. «Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects». In *Metadata and Semantics Research. MTSR 2015. Communications in Computer and Information Science*, a cura di Emmanouel Garaoufallou, Richard J. Hartley, e Panorea Gaitanou, 544:424–436. Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24129-6_37

[7] Daquino, Marilena, Valentina Pasqual, e Francesca Tomasi. «Knowledge Representation of digital Hermeneutics of archival and literary Sources». *JLIS: Italian Journal of Library, Archives and Information Science = Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione: 11, 3, 2020*, fasc. 3 (2020): 59–76. https://doi.org/10.4403/jlis.it-12642

[8] Daquino, Marilena. «Linked Open Data native cataloguing and archival description». *JLIS* 12, fasc. 3 (2021): 91–104. https://doi.org/10.4403/jlis.it-12703

[9] Gangemi, Aldo, Valentina Presutti, Diego Reforgiato, Andrea Giovanni Nuzzolese, Francesco Draicchio, e Misael Mongiovì. «Semantic Web Machine Reading with FRED». *Semantic Web* 8, fasc. 6 (2017): 873–893. https://doi.org/10.3233/SW-160240

[10] Guerrini, Mauro, e Tiziana Possemato. «Linked data: un nuovo alfabeto del web semantico». *Biblioteche oggi* 30, fasc. 3 (2012): 7–15.

[11] Mihindukulasooriya, Nandana, Sanju Tiwari, Carlos F. Enguix, e Kusum Lata. «Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text». In *The Semantic Web – ISWC 2023*, a cura di Terry R. Payne et al., 247–265. Cham: Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-47243-5_14

[12] Polley, Katherine Louise, Vivian Teresa Tompkins, Brendan John Honick, e Jian Qin. «Named Entity Disambiguation for Archival Collections: Metadata, Wikidata, and Linked Data». In *Proceedings of the Association for Information Science and Technology 58*, 1:520–524, 2021. https://doi.org/10.1002/pra2.490

[13] Strötgen, Jannik, e Michael Gertz. «A Baseline Temporal Tagger for All Languages». In *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 541–547. Lisbon, Portugal: Association for Computational Linguistics, 2015. https://doi.org/10.18653/v1/D15-1063

[14] Tomasi, Francesca. «Archival Finding Aids in Linked Open Data between Description and Interpretation». *JLIS.It* 14, fasc. 3 (2023): 134–146. https://doi.org/10.36253/jlis.it-557

[15] Valacchi, Federico. «Not the Institutions but the Subjects Matter. Beyond the Necessary Approximation of Finding Aids?» *JLIS.It* 14, fasc. 3 (2023): 1–14. https://doi.org/10.36253/jlis.it-539.

[16] Valacchi, Federico. «The Parts and the Whole. Integrate Knowledge». *JLIS.It* 13, fasc. 3 (2022): 1–11. https://doi.org/10.36253/jlis.it-477

[17] Vitali, Stefano. «La descrizione degli archivi nell'epoca degli standard e dei sistemi informatici». In *Archivistica. Teorie, metodi, pratiche*, a cura di Linda Giuva e Maria Guercio, 179–210. Roma: Carocci, 2024.

CLEF 2.0. Soluzioni per la catalogazione nativa Linked Data del patrimonio digitale culturale italiano

Marilena Daquino[1], Laurent Fintoni[2], Sebastiano Giacomini[3], Francesca Tomasi[4]
[1] Università di Bologna, Italia - marilena.daquino2@unibo.it
[2] Università di Bologna, Italia - laurent.fintoni2@unibo.it
[3] Università di Bologna, Italia - sebastiano.giacomin2@unibo.it
[4] Università di Bologna, Italia - francesca.tomasi@unibo.it

ABSTRACT
L'affermazione del Web Semantico ha avuto un impatto significativo nel settore delle istituzioni GLAM, per le quali la connessione dei saperi ha assunto una rilevanza tale da produrre numerose iniziative di crowdsourcing e progetti collaborativi di catalogazione nativa Linked Open Data. Una sfida attuale che interessa tali attività collaborative riguarda l'eterogeneità dei contenuti e dei gradi di competenza posseduti dagli utenti. Se da un lato soluzioni esistenti riescono a soddisfare i requisiti minimi in questo ambito di lavoro, spesso a dettare le linee guide dello sviluppo di nuove funzionalità per le applicazioni di crowdsourcing è il concreto impiego di queste stesse piattaforme in contesti di lavoro pratici. Il presente articolo intende analizzare queste esigenze e presentare la soluzione proposta da CLEF 2.0, il software per la catalogazione nativa Linked Open Data adottato in alcuni casi di studio inerenti alla descrizione del patrimonio culturale digitale italiano.

PAROLE CHIAVE
Crowdsourcing; Linked Open Data; catalogazione; patrimonio culturale.

1. INTRODUZIONE
L'affermazione delle tecnologie del Web Semantico ha spinto istituzioni e professionisti a riconsiderare l'organizzazione dei propri saperi e le metodologie per la condivisione della conoscenza. Questa nuova visione del Web, concepito come rete di dati interconnessi (*Linked Data*), ha avuto un impatto significativo sul settore delle istituzioni GLAM (*Galleries, Libraries, Archives, Museums*), offrendo una fondamentale prospettiva di rinnovamento [4] e di superamento del tradizionale isolamento tra le diverse collezioni di dati [7]. Non a caso, l'importanza della condivisione dei saperi tra istituzioni e cittadini ha assunto in tempi recenti una rilevanza tale da produrre una proliferazione di iniziative di *crowdsourcing*, termine col quale riferirsi al contributo di un pubblico non meglio precisato, tramite invito aperto, in attività di varia natura proposte da un'istituzione, organizzazione, o azienda [2].
Nonostante l'enorme potenziale di un approccio collaborativo alla creazione di Linked Data, il crowdsourcing pone tecnici e ricercatori davanti a sfide ancora attuali, come l'eterogeneità dei contenuti e delle competenze degli utenti. Negli ultimi anni, sono emerse varie applicazioni e *content management systems* (CMS) dedicati alla creazione collaborativa di collezioni di Linked Data. Tali soluzioni (cfr. *Sezione 2*) vanno alla ricerca di un sistema di catalogazione omogeneo e coerente, in grado di evitare semplificazioni della realtà e di garantire a tutti gli utenti una facile condivisione dei loro saperi senza limitare le possibilità descrittive dei più esperti. Sebbene alcune di queste applicazioni abbiano saputo far fronte alle richieste di importanti iniziative Linked Data, progetti sempre più complessi e strutturati hanno reso necessarie ulteriori modifiche agli strumenti esistenti. Alla necessità di soddisfare importanti livelli di qualità, usabilità e affidabilità si vanno infatti affiancando nuovi requisiti di completezza, ricchezza e precisione descrittiva.
In questo contesto, il presente contributo intende analizzare la soluzione proposta da CLEF 2.0, la nuova versione dell'applicazione web CLEF[1] (Crowdsourcing Linked Entities via web Form). L'aggiornamento prende le mosse dall'analisi di tre casi di studio: Global Education and Learning (GEL), un progetto co-finanziato dall'UNESCO per la creazione di un catalogo bibliografico sull'educazione alla cittadinanza globale; KNOT, un progetto pilota dell'Università di Bologna e dell'Istituto Centrale per la Digitalizzazione del Patrimonio Culturale del Ministero della Cultura per la valorizzazione del patrimonio culturale digitale degli Atenei italiani; ATLAS, un progetto finanziato dall'Unione europea – Next Generation EU (PRIN2022) per la pubblicazione di un catalogo Linked Open Data di progetti di ricerca Digital Humanities (DH).

[1] https://polifonia-project.github.io/clef/

2. STATO DELL'ARTE

CLEF si inserisce in un contesto già ricco di proposte. Tra gli strumenti di crowdsourcing alternativi, rileva anzitutto Semantic MediaWiki[2], un'estensione aperta e gratuita di MediaWiki, nata con lo scopo di arricchire i tradizionali wiki, costituiti da solo testo, con delle annotazioni semantiche disponibili come Linked Open Data (LOD) nel triplestore del progetto. Alla stessa famiglia appartiene Wikibase[3], progettato per consentire la creazione e gestione collaborativa di Linked Data riutilizzabili in applicazioni esterne, sebbene i suoi dati, pur esportabili in un triplestore Blazegraph[4], siano nativamente conservati in un database relazionale. In ambito GLAM, spicca la piattaforma Omeka S[5], un'evoluzione semantica del software Omeka Classic che fornisce alle istituzioni culturali uno strumento flessibile per la pubblicazione di collezioni di dati, conservati in un database relazionale e unicamente accessibili come documenti JSON-LD tramite API. Diverso è il funzionamento di ResearchSpace[6], un progetto open source curato dal British Museum di Londra, le cui collezioni di Linked Data sono immediatamente disponibili nel triplestore e possono essere interrogate tramite query SPARQL. Chiude questa breve rassegna dei principali CMS, Sinopia[7], un ambiente di editing di Linked Data ottimizzato per la catalogazione di risorse bibliografiche. Tutti i tool descritti si basano sulla creazione di modelli (template), associati a una classe ontologica, per la generazione di schede catalografiche (record o item) tramite web form.

La catalogazione nativa in LOD presenta sfide tecniche e teoriche rilevanti [5]. Un aspetto fondamentale riguarda la cura della qualità dei dati pubblicati, anche alla luce dell'eterogeneità delle fonti. Tra i meccanismi messi a punto per garantire l'affidabilità dei dati condivisi, grande importanza spetta alla documentazione della provenienza (*provenance*) degli asserti. Un punto, questo, spesso ignorato dai tool per la generazione di LOD, ma che appare fondamentale per prevenire incongruenze, porre l'accento sulle responsabilità dei contenuti e rafforzare la fiducia nei dati raccolti [8]. Non a caso, la *provenance* assume un ruolo cruciale anche nell'ambito dei requisiti FAIR[8] (Findability, Accessibility, Interoperability, Reusability), la cui implementazione favorisce la creazione di Linked Data adatti al riuso e all'analisi secondo i metodi di ricerca delle Digital Humanities [6]. Tuttavia, questi stessi principi hanno talvolta assunto un ruolo marginale in progetti di crowdsourcing, dove i dati raccolti sono spesso trattati come dati di serie B e subiscono processi diversi da quelli dei cataloghi ufficiali, portando a scarsa trasparenza e attendibilità.

A questo problema ha tentato di porre rimedio CLEF, una recente proposta software per il crowdsourcing di LOD [3]. CLEF nasce con lo scopo di supportare progetti di piccole e medie dimensioni nella creazione collaborativa di collezioni LOD, facendo ricorso ad uno strumento familiare e di facile utilizzo come quello del Web Form. Attraverso un'interfaccia improntata all'usabilità, l'obiettivo di CLEF è quello di estendere e facilitare l'accesso alla creazione di Linked Data a un bacino di collaboratori sempre più vasto, indipendentemente dalla conoscenza delle tecnologie del Web Semantico. Prioritari in CLEF sono il rispetto dei principi FAIR e la gestione delle informazioni circa la provenienza dei dati. In particolare, il data model definito dalla Provenance Ontology (PROV-O) viene recuperato per registrare le informazioni sul processo editoriale all'interno di *named graphs*. Queste informazioni, insieme ai dati delle risorse catalogate, sono rese immediatamente disponibili tramite triplestore (Blazegraph) e interrogabili mediante un apposito endpoint SPARQL, diversamente da altri CMS basati su database relazionali.

3. RACCOLTA DEI REQUISITI E ANALISI DEI COMPETITOR

Tra le recenti attività di ricerca che hanno spinto verso un aggiornamento dei sistemi per il crowdsourcing di LOD, figurano i seguenti casi di studio: GEL, un progetto concluso, volto alla produzione e all'aggiornamento di una *knowledge base* bibliografica sulla letteratura circa i temi dell'educazione alla cittadinanza globale; KNOT, un progetto per la valorizzazione del patrimonio culturale digitale degli Atenei italiani e, anzitutto, dei progetti e dei prodotti della ricerca in ambito DH; ATLAS, avviato a ottobre 2023 per il censimento e l'estrazione di dati a partire da progetti di DH incentrati sul patrimonio culturale italiano. Fondamentale è la definizione di un workflow per la raccolta, normalizzazione, bonifica e trasformazione dei dati in formato RDF al fine di integrare tale processo nella quotidiana attività editoriale di catalogazione.

Da un'analisi preliminare e dal confronto con i committenti dei progetti, sono emersi i seguenti requisiti:

- **Entity Linking**. Fondamentale nella produzione di LOD è la creazione di collegamenti con risorse ed entità presenti sul web per mezzo dei loro URI [9]. Tuttavia, individuare le esatte entità di interesse può risultare

[2] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki
[3] https://wikiba.se/
[4] https://blazegraph.com/
[5] https://omeka.org/s/
[6] https://researchspace.org/
[7] https://sinopia.io/
[8] https://www.go-fair.org/fair-principles/

complesso, specialmente per utenti poco esperti messi a confronto con enormi moli di dati. Appare dunque indispensabile potenziare gli strumenti che assistono gli utenti nel recupero di entità online, fornendo suggerimenti automatici in tempo reale e metodi di disambiguazione.

- **Datatype temporali**. Le complessità descrittive delle risorse catalogate nei casi di studi menzionati evidenziano la necessità di ampliare il numero di tipologie di dati (*datatype*) fruibili da parte dell'utente finale, specialmente quando si ha a che fare con date e periodi dai confini incerti.
- **Vocabolari controllati e thesauri SKOS**. Una funzionalità di grande utilità consiste nel recupero e riuso di nomenclature standardizzate come nel caso dei vocabolari SKOS, delle risorse controllate i cui termini, dal significato comune e condiviso, risultano spesso fondamentali ai fini di una descrizione puntuale.
- **Multimedia**. Le possibilità descrittive di un'entità non si limitano ai soli dati testuali ma possono coinvolgere anche altri media come immagini, video, file audio e anteprime di pagine web esterne. L'introduzione di una preview delle risorse in rete descritte (tramite inclusione di un *iframe*) consentirebbe di completare la descrizione dell'oggetto in esame senza richiedere al visitatore del catalogo di abbandonare la pagina web in cui sono citate.
- **Knowledge Extraction**. Un ultimo strumento dovrebbe garantire la possibilità di supportare l'estrazione semiautomatica di entità chiave per la descrizione di un'entità a partire da una fonte di dati online (e.g. SPARQL endpoint, file statico, API). Questa funzionalità può prendere la forma di servizi di Named Entity Recognition da un testo pieno importato nella piattaforma di catalogazione, o può essere disegnata da un utente esperto proponendo una query personalizzata, basata sulle caratteristiche della fonte da interrogare.

Sebbene alcuni tra i CMS precedentemente menzionati siano già in grado di soddisfare le richieste in esame, a mancare è un'integrazione coerente e di facile utilizzo delle stesse. Dall'analisi dei sistemi esistenti emerge infatti il seguente quadro.

Entity Linking. In CLEF, questo genere di funzionalità è già presente, seppure le entità siano al momento estraibili soltanto da Wikidata e dal catalogo stesso. Similmente, Sinopia e Omeka S consentono agli utenti di ricevere dei suggerimenti da delle basi di dati predefinite, mentre in ResearchSpace i creatori di Template possono specificare una query SPARQL per generare automaticamente dei suggerimenti. Semantic MediaWiki, consente il riuso di entità da Wikidata e di dati estratti tramite API costruite *ad hoc* o attraverso un'apposita espansione.

Datatype temporali. All'interno di CLEF, le opzioni disponibili sono limitate a Literal (stringhe) e URI, escludendo di fatto altri tipi di dati fondamentali come date e periodi temporali. La definizione di datazioni è disponibile in Sinopia, mentre in Omeka S è possibile utilizzare un apposito modulo per la gestione di datatype numerici, tra cui anche le date, che consente di associare ad una voce di un Template un triplice campo di input (Anno, Mese, Giorno) ed un timespan opzionale. A seconda dei dati forniti, il sistema salverà la datazione in input associandola al datatype più appropriato. Segue un approccio simile anche Semantic MediaWiki, mentre molto più ricca è la scelta in ResearchSpace, dove a ciascun campo è associabile uno fra i datatype specificati in RDF 1.1.

Vocabolari controllati e thesauri SKOS. L'utilizzo di vocabolari controllati è una funzionalità integrata in Omeka S in seguito all'installazione del relativo modulo. Oltre a una ricca scelta iniziale, gli utenti possono introdurre dei nuovi vocabolari specificandone manualmente i termini. Similmente, in CLEF, gli utenti accreditati possono nativamente inserire liste di termini controllate nella forma <label, URI>. Tuttavia, importare liste di dimensioni più ampie può essere fonte di errori e ogni campo può essere associato ad un solo vocabolario. Sinopia offre un'ampia scelta di risorse controllate, ma non prevede alcuna funzione per l'integrazione di nuovi vocabolari, mentre Semantic MediaWiki e ResearchSpace si avvalgono degli strumenti già descritti per l'Entity Linking.

Multimedia. L'utilizzo di risorse multimediali è presente in tutti i CMS tranne Sinopia e CLEF. Sistemi come Omeka S e ResearchSpace consentono l'integrazione di strumenti più avanzati come un visualizzatore IIIF. Meno frequente è invece il ricorso agli iframe, disponibili comunque in Omeka S e in Semantic MediaWiki.

Knowledge Extraction. Al fine di operare un recupero di ampie moli di dati, sarebbe auspicabile fornire ai collaboratori più esperti uno strumento per l'estrazione semiautomatica di dati tramite interrogazioni mirate. Attualmente, nessuno dei sistemi esaminati offre questo tipo di servizio, la cui implementazione dovrebbe andare oltre al solo recupero di dati RDF, includendo anche l'estrazione di informazioni da altre fonti (e.g. API e file statici).

Rispetto ai suoi competitor, CLEF mostra alcuni punti di attenzione che sono al vaglio. Non di meno, l'assenza totale di soluzioni per knowledge extraction semi-automatica consente di valutare positivamente la possibilità di espandere CLEF rispetto alle altre soluzioni data la modularità e sostenibilità del codice e la possibilità di effettuare tali integrazioni in un sistema che sia effettivamente nativo LOD. Un ulteriore obiettivo è quello di comprendere come ottimizzare i meccanismi e le funzionalità esistenti all'interno di un sistema unificato e coerente. In tal senso, CLEF ha dimostrato la sua capacità di fornire un ambiente di lavoro favorevole all'integrazione di nuovi strumenti, preservando le sue caratteristiche originali.

4. CLEF: IL SISTEMA COLLABORATIVO DI CATALOGAZIONE NATIVO LOD

Il funzionamento di CLEF si basa sulla suddivisione dei ruoli tra membri accreditati e contributori. Ai primi spetta il compito di definire i parametri fondamentali del progetto in costruzione (e.g. nome del catalogo, l'endpoint), gestire l'organizzazione dei dati attraverso la definizione di Template e validare i dati inseriti per la pubblicazione. Più di preciso, la creazione di un Template consiste nella combinazione e definizione di molteplici campi di input tra quelli messi a disposizione dalla piattaforma, che nella versione originaria di CLEF include le seguenti opzioni:

- Textbox. Un campo di testo, destinato ad accogliere alternativamente una fra le seguenti tipologie di valori: brevi stringhe di testo, stringhe riconciliate ad URI di entità provenienti da Wikidata e dal catalogo stesso, stringhe riconciliate a località geografiche di GeoNames. Questa funzionalità è presente anche in Omeka S.
- Textarea. Un campo di input per descrizioni testuali estese, dalle quali è possibile estrarre entità nominate riconciliate a Wikidata, tramite un sistema di Named Entity Recognition e Data Reconciliation. Funzione non disponibile in nessuno dei software analizzati.
- Dropdown. Menù a tendina popolato con una serie di etichette associate a URI, tra cui l'utente finale sarà chiamato a selezionare un solo valore. Questa funzionalità si ritrova anche negli altri competitor di CLEF.
- Checkbox. Segue un funzionamento del tutto analogo a quello di un Dropdown pur consentendo all'utente finale di selezionare più di un valore tra quelli proposti. Questa funzionalità non è presente in altri competitor.

CLEF 2.0 interviene su questa lista di opzioni per arricchire il sistema con funzionalità che rispondono ai requisiti raccolti. Ad essere aggiornati non sono soltanto le possibilità di interazione degli utenti, ma anche il sistema di gestione che consente la generazione, serializzazione e salvataggio di Linked Data, oltre alla visualizzazione finale dei dati.

Prima di procedere con l'illustrazione delle nuove funzionalità proposte, occorre tuttavia notare come solo alcune delle sfide affrontate abbiano potuto trovare una risposta efficace senza presuppore un livello minimo di conoscenza e consapevolezza, in materia di tecnologie semantiche, da parte dei creatori di un progetto CLEF. Se lo scopo è infatti quello di garantire un riuso quanto più ampio e libero possibile di servizi e risorse già presenti in rete, fondamentale dovrà essere l'apporto dei membri del nuovo progetto nella configurazione delle funzionalità offerte da CLEF, vale a dire nella fase di definizione dei *templates*. È ad esempio il caso del recupero di termini controllati da *thesauri* SKOS: un'opzione, questa, che viene spesso limitata dai software attuali a una serie predefinita di basi di dati, garantendo, il più delle volte, soltanto aggiunte manuali. L'estensione dell'utilizzo di una simile funzionalità a risorse nuove non può dunque prescindere dalla specificazione di alcune chiavi di accesso (e.g.: endpoint SPARQL, query SPARQL di default) al servizio desiderato, ad opera dei membri accreditati del catalogo in costruzione. Spetta poi all'applicazione di base il compito di guidare gli utenti nella configurazione dei parametri richiesti, così come quello di assicurare una corretta gestione delle interazioni con i servizi selezionati.

A dispetto di queste complessità, resta fondamentale nell'aggiornamento a CLEF 2.0, l'intento di preservare la possibilità di un utilizzo basilare delle funzionalità dell'applicazione, così da garantire la creazione di un catalogo strutturato e completo anche in caso di totale assenza di competenze tecniche. Pertanto, l'applicazione è stata aggiornata come segue.

Entity Linking. CLEF integra già una funzionalità di suggerimento automatico basata sul recupero di entità da Wikidata e dal catalogo stesso. Sebbene Wikidata fornisca una base di conoscenza estremamente ricca, altre risorse semantiche offrono maggiore copertura in settori più vicini al patrimonio culturale. Tra queste, VIAF[9] (Virtual International Authority File), offre uno strumento fondamentale per l'identificazione di entità legate all'universo bibliografico. Per questo motivo, CLEF 2.0 integra VIAF nella lista di suggerimenti qualora una stringa di input non dovesse restituire risultati in Wikidata. Nel caso di studio KNOT, risulta fondamentale anche la possibilità di menzionare l'URL di risorse web collegate, da gestire separatamente dalle entità recuperate da Wikidata, VIAF, catalogo e GeoNames. Per supportare questo nuovo tipo di input, CLEF 2.0 integra un *value type* da associare a textbox in fase di creazione di un Template. Tutte le risorse selezionate sono immediatamente raggiungibile dalla scheda catalografica di riferimento.

Datatype temporali. Nella sua prima versione, CLEF prevede solo l'input di URI e valori aventi datatype xsd:string. Per far fronte ad esigenze di maggiore accuratezza descrittiva, l'aggiornamento introduce tre nuovi datatype per le datazioni: xsd:Date (YYYY/MM/DD), xsd:gYearMonth (YYYY/MM) e xsd:gYear (YYYY).

Vocabolari controllati e thesauri SKOS. CLEF 2.0 introduce la possibilità di creare campi per l'inserimento di termini da vocabolari controllati. È possibile selezionare il vocabolario di riferimento da una lista limitata di risorse già integrate[10], fondamentali per la descrizione di progetti di Digital Humanities, o integrarne degli altri mediante query SPARQL ad un

[9] https://viaf.org/viaf/

[10] Si tratta del Vocabolario delle Licenze (https://schema.gov.it/lodview/controlled-vocabulary/licences), di alcune authority tables prodotte dall'Unione Europea (https://op.europa.eu/en/web/eu-vocabularies/authority-tables) e della Taxonomy of Digital Research Activities in the Humanities (TaDiRAH, https://vocabs.dariah.eu/tadirah/en/).

servizio online. Nel secondo caso, durante la definizione del template si inserisce l'URL della risorsa, l'endpoint e una query SPARQL per estrarre termini sulla base di una stringa di input. In fase di inserimento dati, l'utente vedrà una lista di suggerimenti automatici analoghi a quelli del campo textbox. Per garantire una visuale completa delle risorse terminologiche disponibili, agli utenti vengono forniti dei collegamenti rapidi ai vocabolari selezionati, agevolandone la consultazione.

Multimedia. CLEF 2.0 introduce due nuove funzionalità (Multimedia e Preview) per l'inserimento di riferimenti URL a file multimediali presenti in rete (immagini, audio, video e preview di pagine web esterne). Poiché CLEF si propone come strumento agile per la catalogazione nativa LOD e non per la creazione di cataloghi multimediali, non è possibile importare file, ma è necessario garantirne la conservazione altrove. Per ovviare a tale limitazione, CLEF consente di selezionare quali URL si riferiscono ad una risorsa digitale online per la quale non si ha a disposizione una strategia di preservazione a lungo termine e provvede a inviare una richiesta di deposito di uno snapshot della risorsa web in questione su Internet Archive[11].

Knowledge extraction. La creazione di Linked Data si rivela un processo dispendioso quando l'inserimento dei valori va svolto manualmente. I tool esistenti si sono occupati solo parzialmente di offrire un vero e proprio strumento per l'estrazione automatica di dati. Pertanto, CLEF 2.0 introduce un prototipo di Knowledge Extraction, progettato per l'estrazione automatica di entità da una varietà di fonti online: API, endpoint SPARQL e file statici in formato JSON e CSV. Questa funzionalità ha richiesto lo sviluppo di funzioni anche complesse per gestire correttamente le possibili casistiche. Al momento, si è scelto di limitare l'utilizzo di questo strumento all'estrazione di liste di URI e relative label, utilizzando una proprietà di default (*schema:keywords*) per relazionare il record alle entità estratte. L'interrogazione a un servizio API richiede tre input da parte dell'utente: l'URL dell'API, i parametri della richiesta AJAX e il percorso per identificare URI ed etichette nella risposta (JSON) fornita dall'API. Più semplice è l'interrogazione di un endpoint SPARQL, che necessita soltanto di un URL e di una query da eseguire. Per interrogare i file statici, CLEF 2.0 ricorre a SPARQL Anything [1], il cui funzionamento è quello di un tipico endpoint SPARQL, con la particolarità di poter generare dati RDF a partire da una vasta gamma di fonti di dati non nativamente RDF. Si richiedono l'URL della risorsa desiderata e una query SPARQL. Una volta estratti dei risultati, la lista di URI e label viene proposta all'utente, che può curare manualmente e scegliere se eliminare alcune istanze. Ulteriori estrazioni possono essere effettuate, anche selezionando fonti diverse. Per ciascuna estrazione effettuata, il catalogo crea un nuovo grafo RDF in cui conservare i risultati ottenuti, l'autore e il metodo utilizzato per l'estrazione. Lo stesso grafo viene poi collegato a quello del record.

5. DISCUSSIONE E CONCLUSIONE

Il presente contributo ha esaminato l'attuale panorama della creazione di collezioni di LOD, evidenziando l'esigenza di rinnovare i sistemi esistenti secondo nuovi requisiti. Dall'analisi dei principali sistemi per la catalogazione collaborativa di Linked Data, sono emersi i punti di forza di CLEF, l'applicazione web scelta come punto di partenza del presente lavoro. Sulla base dei casi di studio KNOT, ATLAS e GEL sono stati definiti i nuovi requisiti per l'aggiornamento di CLEF, con brevi analisi delle soluzioni presenti in altri CMS per ciascuna delle nuove funzioni richieste. Si è dato quindi risalto all'esigenza di rafforzare gli strumenti per il riutilizzo di entità estratte da basi di dati strutturate (VIAF, thesauri SKOS) e non strutturate (SPARQL Anything), e alla necessità di ampliare le tipologie di dati e le risorse multimediali disponibili.

La novità principale di CLEF 2.0 consiste nell'integrazione di uno strumento per la Knowledge Extraction: una funzionalità avanzata, che cerca di soddisfare esigenze più vicine ad un pubblico esperto, come quello che si accinge a catalogare i progetti di Digital Humanities relativi al patrimonio culturale italiano (KNOT e ATLAS). Non di meno, la possibilità di estrarre dati da file statici rappresenta il miglior compromesso per consentire ad utenti meno esperti di lavorare su risorse meno tecniche. Sebbene l'attuale implementazione richieda familiarità con SPARQL per estrarre i dati, in apparente contrasto con gli obiettivi di usabilità di CLEF, i benefici derivanti dal ricorso a SPARQL Anything rimangono preponderanti. In primo luogo, l'utilizzo di uno strumento unico per l'analisi di più formati di file sostituisce lo sviluppo *ad hoc* di nuove soluzioni per ogni nuovo formato inserito. Al tempo stesso, l'impiego di un motore di query SPARQL trasferisce sull'analisi di file statici tutti i vantaggi di un linguaggio di interrogazione, incluso un efficiente meccanismo per il filtraggio dei dati, fondamentale in presenza di documenti di grandi dimensioni. Inoltre, SPARQL Anything richiede ai propri utenti la conoscenza di un unico linguaggio per l'interrogazione, semplificando i requisiti di conoscenza pregressa. Ulteriori complessità tecniche potrebbero interessare, come anticipato, gli utenti accreditati di un progetto CLEF, ai quali spetta il compito di definire i campi di input all'interno di ciascun Template. Tra le nuove funzionalità proposte, CLEF 2.0 ha introdotto la possibilità di indicare risorse esterne per il recupero e riuso di termini controllati (Vocabolari controllati e thesauri SKOS). I parametri richiesti per l'impiego di tale soluzione includono l'endpoint SPARQL della risorsa a cui si intende accedere e una query SPARQL per estrarre i termini controllati. Sulla base di questi verranno successivamente

[11] https://archive.org/

generati dei suggerimenti automatici per l'utente finale. Sebbene la specificazione di tali parametri potrebbe risultare complessa per utenti meno esperti, CLEF continua ad offrire alternative non automatiche più semplici da utilizzare. È il caso, ad esempio, dei campi di input di tipo Checkbox e Dropdown, che consentono ai creatori del Template di specificare manualmente i termini desiderati, garantendo di fatto lo stesso risultato finale, pur senza alcun meccanismo di estrazione automatica e di autocompletamento.

Non richiedono invece alcuna competenza in materia di tecnologie semantiche le restanti funzionalità proposte, incluse l'estrazione automatica di entità da VIAF e l'introduzione di *datatype* temporali.

Un ultimo aspetto cruciale dell'aggiornamento di CLEF 2.0 riguarda il ricorso a risorse multimediali online. Spesso, gli utenti non gestiscono direttamente i documenti e i servizi richiesti, sollevando questioni sulla conservazione online delle risorse. Appare dunque logico ribadire l'importanza riservata da CLEF ad una strategia trasversale per la preservazione a lungo termine di risorse online.

Nonostante alcune criticità evidenziate riguardo allo sviluppo di soluzioni per Knowledge Extraction, l'insieme delle sfide, dei traguardi e delle limitazioni discusse in questo contributo prepara il terreno per ulteriori analisi e sviluppi futuri. L'obiettivo primario rimane la realizzazione di un'applicazione sempre più completa e intuitiva, al fine di rendere i Linked Open Data accessibili a tutti.

6. RINGRAZIAMENTI

Questo progetto ha ricevuto finanziamento dall'Unione europea – Next Generation EU (PRIN2022, 20227M8RS7) e dal programma European Union's Horizon 2020 research and innovation (GA 101004746).

BIBLIOGRAFIA

[1] Daga Enrico, Luigi Asprino, Paul Mulholland, e Aldo Gangemi. «Facade-X: An Opinionated Approach to SPARQL Any- thing». In *Studies on the Semantic Web*, a cura di Mehwish Alam, Paul Groth, Victor De Boer, Tassillo Pellegrini, Harshvardhan J. Pandit, Elena Montiel, Víctor Rodríguez Doncel, Barbara McGillivray, e Albert Meroño-Peñuela, 58–73. 53. IOS Press, 2021. https://doi.org/10.3233/SSW210035

[2] Daquino, Marilena. «Linked Open Data native cataloguing and archival description». *JLIS* 12, fasc. 3 (2021): 91–104. https://doi.org/10.4403/jlis.it-12703

[3] Daquino, Marilena, Mari Wigham, Enrico Daga, Lucia Giagnolini, e Francesca Tomasi. «CLEF. A Linked Open Data native system for Crowdsourcing». *arXiv*, 2022. https://doi.org/10.48550/arXiv.2206.08259

[4] Davis, Edie, e Bahareh Heravi. «Linked Data and Cultural Heritage: A Systematic Review of Participation, Collaboration, and Motivation». *Journal on Computing and Cultural Heritage* 14, fasc. 2 (2021): 1–18. https://doi.org/10.1145/3429458

[5] Hawkins, Ashleigh. «Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-Digital Archives via the Semantic Web». *Archival Science* 22, fasc. 3 (2022): 319–344. https://doi.org/10.1007/s10502-021-09381-0

[6] Hermon, Sorin, e Franco Nicolucci. «FAIR Data and Cultural Heritage Special Issue Editorial Note». *International Journal on Digital Libraries* 22, fasc. 3 (2021): 251–255. https://doi.org/10.1007/s00799-021-00309-8

[7] Marden, Julia, Carolyn Li-Madeo, Noreen Whysel, e Jeffrey Edelstein. «Linked Open Data for Cultural Heritage: Evolution of an Information Technology». In *Proceedings of the 31st ACM International Conference on Design of Communication*, 107–12. Greenville North Carolina USA: ACM, 2013. https://doi.org/10.1145/2507065.2507103

[8] O'Hara, Kieron, Harith Alani, Yannis Kalfoglou, e Nigel Shadbolt. «Trust strategies for the semantic web». In *Proceedings of the 2004 International Conference on Trust, Security, and Reputation on the Semantic Web. ISWC'04*, 127:42–51. Aachen, DEU: CEUR-WS.org, 2004.

[9] Thanos, Costantino. «Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies». *Publications*, 2, 5, fasc. 1 (2017). https://doi.org/10.3390/publications5010002

L'ontologia BiGraFo: verso un modello semantico per l'opera di Franco Fortini

Laura Antonietti[1], Emilio M. Sanfilippo[2], Emmanuela Carbé[3]
[1] Università degli Studi di Siena, Italia – laura.antonietti2@unisi.it
[2] CNR ISTC Laboratorio di ontologia applicata, Italia – emilio.sanfilippo@cnr.it
[3] Università degli Studi di Siena, Italia – emmanuela.carbe@unisi.it

ABSTRACT
Il contributo propone i primi risultati del progetto BiGraFo, avviato nel settembre 2023 per realizzare un catalogo semantico dedicato all'opera di Franco Fortini. Ci si concentra in particolare sugli aspetti metodologici ed euristici inerenti alla riflessione nata intorno alle ontologie bibliografiche e alla definizione di un modello per organizzare e rappresentare un caso particolarmente complesso come quello di Fortini: la sua ingente produzione e il continuo rimaneggiamento dei testi, spesso riproposti in sedi editoriali differenti, rendono l'autore un banco di prova significativo nell'ambito della rappresentazione della conoscenza.

PAROLE CHIAVE
Digital bibliography; bibliographic ontologies; Franco Fortini.

1. INTRODUZIONE
Il progetto BiGraFo nasce per mappare e indagare la produzione intellettuale e artistica di Franco Fortini (1917-1994), poeta, traduttore e critico letterario, il cui patrimonio documentario – archivio e biblioteca d'autore – è oggi conservato presso la Biblioteca di Area Umanistica dell'Università di Siena, dove Fortini insegnò tra il 1971 e il 1989: si tratta di un fondo di oltre 5000 lettere e 300 autografi, ma anche disegni, incisioni e dipinti, e collezioni speciali di materiali contenuti in supporti fragili come audiocassette e Floppy Disk, oggi in fase di recupero. All'archivio si aggiunge un fondo librario che consta di circa 5000 volumi – molti dei quali postillati – appartenuti a Fortini.
Data l'importanza del materiale conservato, il gruppo di lavoro ha ritenuto opportuno dare avvio a un progetto di studio e valorizzazione che includa anche l'impiego di metodi e strumenti digitali, laddove risultino utili per fornire l'accesso e la disseminazione del patrimonio materiale e immateriale: si è deciso di partire dal nucleo centrale di ogni ricerca, la bibliografia, valorizzando il lavoro già in atto da molti anni grazie ad archivisti, bibliotecari e ricercatori che collaborano all'archivio di Franco Fortini.
L'obiettivo del progetto è la pubblicazione di una piattaforma digitale, il cui strumento più importante è un catalogo aperto e incrementabile per interrogare l'opera dell'autore, di cui è stata recentemente pubblicata la *Bibliografia degli scritti di Franco Fortini* [2]. Basato su un'ontologia nei linguaggi del Semantic Web, il catalogo sarà predisposto ad accogliere e mettere in relazione anche la bibliografia della critica, nonché a dare rilievo al patrimonio dell'archivio e della biblioteca d'autore; la piattaforma includerà infine una mostra virtuale dedicata a Fortini e alla sua opera, fornendo un percorso espositivo di immagini e testi per valorizzarne la sua figura.
Il progetto si inserisce all'interno del contesto più ampio della rappresentazione della conoscenza e della gestione dei dati d'autore in ambito bibliografico (e, in prospettiva, archivistico): la riflessione sul metodo e sull'apporto euristico dell'attività di modellizzazione costituisce uno degli aspetti più interessanti e rilevanti, su cui in questa sede intendiamo focalizzare la nostra attenzione.

2. STATO DELL'ARTE E METODOLOGIA
A questo stadio iniziale del progetto la nostra riflessione si concentra sugli aspetti metodologici e sulla definizione del modello ontologico, affiancato da un corretto processo di documentazione. L'analisi di progetti simili e delle ontologie bibliografiche esistenti hanno costituito i primi passi fondamentali in una fase preliminare del lavoro.
Nel mondo degli archivi e delle biblioteche gli esempi di applicazione di modelli ontologici e di pubblicazione di *Linked Open Data* sono numerosissimi e a più livelli[1]; tuttavia non si riscontrano, a nostra conoscenza, progetti italiani che adottino modelli orientati alle ontologie per una cura specifica e aggiornata del dato bibliografico d'autore (inteso nel senso più

[1] Si pensi, in ambito nazionale, alla piattaforma dati.beniculturali.it del Mibac, e al Knowledge Graph di ArCo – The Italian Cultural Heritage Knowledge Graph (http://wit.istc.cnr.it/arco/?lang=en), ma anche al soggettario della Biblioteca Nazionale di Firenze [11], o ancora a singoli progetti come il catalogo dell'archivio della Fondazione Federico Zeri di Bologna (http://data.fondazionezeri.unibo.it).

ampio), che qui si intende strutturare – con possibili aperture al dato archivistico – in un ambiente in grado di attuare un modello di rappresentazione della conoscenza [19], raggiungendo in prospettiva territori di esplorazione filologica e critica [13].

Sono state prese in esame le principali ontologie bibliografiche, in particolare BIBO, BIBFRAME e FaBiO[2]. L'analisi comparativa dei tre modelli [1, 4, 5, 9, 12] e l'esame di una campione significativo del dataset bibliografico dell'opera di Fortini hanno dimostrato che il progetto può orientarsi, con integrazioni di cui si dirà a breve, verso l'ontologia FaBiO - *The FRBR-aligned Bibliographic Ontology* [14, 16][3].

FaBiO è un'ontologia sviluppata sulla base del modello FRBR, in particolare sulla versione *core*[4], di cui eredita la struttura concettuale e formale. Ad alto livello le sue classi principali sono quindi quelle di *Work*, *Expression*, *Manifestation* e *Item,* le quali costituiscono la cosiddetta struttura *WEMI* di FRBR. FaBiO estende FRBR attraverso vari elementi di modellazione (classi e relazioni) per facilitare la rappresentazione dei dati in ambito bibliografico. Tra gli elementi che lo distinguono da FRBR possiamo menzionare le relazioni che permettono di collegare in modo diretto le istanze di *Work* e *Manifestation* (*has manifestation*), *Work* e *Item* (*has portrayal*), o ancora *Expression* e *Item* (*has representation*) senza necessariamente passare per le classi (o istanze) intermedie (vd. Fig. 1): stabilendo delle relazioni dirette tra le classi (che potremmo definire *shortcut* sintattiche), FaBiO agevola l'accesso e l'interrogazione ai dati.

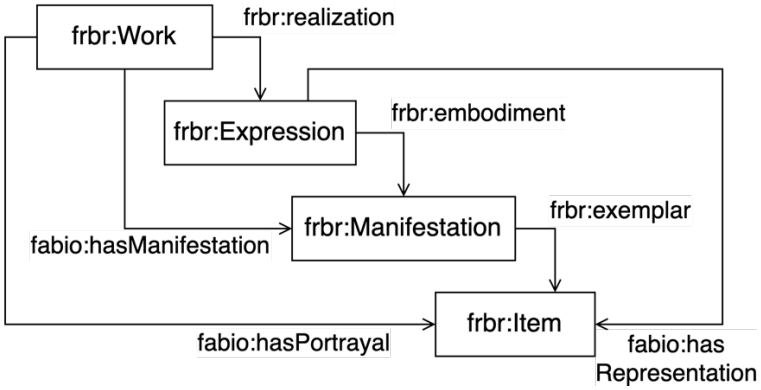


Figura 1. Classi WEMI di FRBR con le shortcut *di FaBiO*

Le qualità intrinseche di FaBiO non sono esenti da qualche criticità nel contesto della modellizzazione di alcune sottoclassi rispetto alla struttura di FRBR [4, 5]; va inoltre osservato che BIBO e BIBFRAME sono due modelli maggiormente adottati a livello progettuale. FaBiO ha tuttavia alcune caratteristiche particolarmente adatte a BiGraFo:

1. Una maggiore espressività e granularità, essenziali per rappresentare le complessità proprie dei dati bibliografici d'autore, come ad esempio le frequenti ripubblicazioni in sedi editoriali differenti (è il caso di un testo, saggistico o poetico, uscito su rivista e confluito in un'opera a stampa) e il complesso quadro di varianti d'autore nelle diverse edizioni. Gli elementi di FaBiO, e in particolar modo la struttura *WEMI,* risultano in questo contesto particolarmente adatti.
2. La possibilità di integrare facilmente informazioni relative ai diversi ruoli implicati nel processo di pubblicazione editoriale, fondamentale per la rappresentazione nel modello delle opere postume e utile anche nella prospettiva della rappresentazione della bibliografia della critica. FaBiO è una delle ontologie SPAR, *Semantic Publishing and Referencing Ontologies* [15], una *suite* di ontologie complementari che coprono molteplici aspetti della pubblicazione e della descrizione bibliografica: saranno di particolare interesse per il nostro progetto, oltre a FaBiO, anche le ontologie PRO (*Publishing Roles Ontology*) e SCoRO (*Scholarly Contributions and Roles Ontology*).
3. L'opportunità di relazionare i dati bibliografici con quelli relativi ai materiali d'archivio: pur non escludendo di integrare in futuro un modello ontologico specifico per una descrizione archivistica strutturata e dettagliata del Fondo Fortini, allo stato attuale FaBiO ci consente, tramite alcune delle sue classi quali ad esempio *Archival record* (e

[2] Bibliographic Ontology (BIBO): https://www.dublincore.org/specifications/bibo/ Bibliographic Framework Initiative (BIBFRAME): https://www.loc.gov/bibframe/; e FRBR-aligned Bibliographic Ontology (FaBiO), http://www.sparontologies.net/ontologies/fabio.

[3] BIBFRAME, come FaBiO, integra il modello FRBR: se FaBiO ne mantiene la struttura quadripartita (*Work, Expression, Manifestation, Item*), BIBFRAME accorpa *Work* e *Expression* in un'unica classe. Per il confronto e l'allineamento con FRBR, si vedano ancora i saggi di Biagetti [4, 5]. Vale la pena ricordare che l'ontologia FaBiO non è attualmente allineata con la versione corrente di LRMoo (LRM object-oriented: https://cidoc-crm.org/frbroo/sites/default/files/LRMoo_V0.9.6.pdf), evoluzione del modello FRBRoo.

[4] Expression of Core FRBR Concepts in RDF: https://vocab.org/frbr/core.

Archival record set) a livello di *Work*, e *Archival document* (e *Archival document set*) a livello di *Expression*, di raggiungere dei primi obiettivi[5].

In sintesi, l'ontologia FaBiO presenta una struttura che, estendendo quella di FRBR, facilita l'uso di quest'ultima in contesti applicativi concreti: l'estensione ne eredita i vantaggi ma anche le problematicità, in particolar modo, da un punto di vista concettuale, quelle relative alla nozione di *Work* [8, 17]. Nell'ultima versione del modello LRMoo, a proposito della classe *Work,* leggiamo: «This class comprises distinct intellectual ideas conveyed in artistic and intellectual creations, such a poems, stories or musical compositions. […] The main purpose of this class is to enable bringing together intellectually equivalent Expressions in order to display to a user all available alternatives of the same intellectual or artistic content» [3]. La caratterizzazione di *Work* rispetto alle "idee" (*intellectual ideas*) e al "contenuto" di testi intellettualmente equivalenti (*intellectually equivalent Expressions*) ha creato non poche ambiguità nella ricezione del modello stesso: la classe *Work* viene infatti intesa in alcuni casi in riferimento alla volontà dell'autore [10], in altri come elemento formale utile alla rappresentazione di più testi simili sul profilo dei loro contenuti [7].

Limitandoci a considerare solo queste due interpretazioni, da un punto di vista ontologico si tratta di due entità ben distinte: nel primo caso, potremmo parlare di *Work* in senso *autoriale*, l'idea dell'opera così come intesa dall'autore; nel secondo di *Work* in senso *documentale*, prendendo in prestito la terminologia di Smiraglia [18], come elemento di modellazione utilizzato in un sistema informativo per classificare più testi. La prima lettura pone l'annoso problema empirico dell'*intentio auctoris*, per non menzionare il dibattito di barthesiana memoria sulla rilevanza dell'autore nell'interpretazione della sua opera. La seconda lettura, di derivazione biblioteconomica, offre per certi versi una visione più snella e coerente rispetto alla prima: la classe *Work* è introdotta per organizzare più testi indipendentemente dall'intenzionalità dell'autore. In BiGraFo si adotta questa seconda lettura, non senza riscontrare qualche ambiguità: diversi testi di Fortini possono ad esempio essere pensati come alternative di una stessa classe *Work*, ma nel contempo rappresentare autonomamente *Work* differenti.

3. MODELLO E CASI STUDIO

Il modello adottato per descrivere la produzione fortiniana, come quello di FaBiO, è stato strutturato sulla scorta delle categorie *WEMI* di FRBR. La rappresentazione di *Item* come oggetti fisici è stata per il momento messa da parte, senza escludere tuttavia la possibilità di integrare in futuro i dati relativi alle collezioni librarie presenti nella Biblioteca di Area Umanistica di Siena. Le classi *Work, Expression* e *Manifestation* sono invece risultate particolarmente adatte alla descrizione della complessa produzione dell'autore, caratterizzata dalle frequenti ripubblicazioni delle opere in sedi editoriali differenti e, ancora, dal continuo rimaneggiamento delle stesse.

Per meglio illustrare come il modello di FaBiO sia stato adattato e integrato al progetto, portiamo a titolo d'esempio il caso di *Al di là della speranza*, un testo poetico che Fortini compone nel 1956. La poesia è stata trasmessa nel corso degli anni in tre differenti versioni testuali: una versione su rivista, più breve, una versione integrale e una versione decurtata, privata dall'autore delle prime strofe. Le ultime due versioni hanno conosciuto diverse edizioni: il testo integrale è stato pubblicato la prima volta sulla rivista «Officina» nel 1957 e ripreso successivamente in volume nel 1959 (in *Poesia ed errore*, Feltrinelli) e ancora nel 1987 (in *Versi primi e distanti 1937-1957,* All'Insegna del pesce d'oro). Il testo decurtato è uscito invece la prima volta nel 1969 nell'edizione mondadoriana di *Poesia e errore* e successivamente nel 1974 nel volume di *Poesie scelte (1939-1973)* a cura di Pier Vincenzo Mengaldo; segue un'edizione nel 1978 (in *Una volta per sempre*, Einaudi) e nel 1990 (in *Versi scelti. 1939-1989*, Einaudi).

Dal punto di vista del modello (vd. Fig. 2)[6], si è deciso di distinguere tre diverse istanze di *Expression* – una per l'anticipazione su rivista, una per la versione integrale e un'altra per la versione decurtata – con le relative edizioni rappresentate come *Manifestation*, tutte raggruppate da un unico *Work*. Il modello di FaBiO risulta dunque particolarmente adatto alla rappresentazione concettuale della complessità del dato d'autore e lo è anche dal punto di vista funzionale della sua interrogazione: poter identificare a quale versione autoriale si rifanno le diverse edizioni è di estremo interesse nella prospettiva di un'indagine filologica più approfondita.

[5] Si può pensare, in prospettiva, all'impiego dell'ontologia CIDOC-CRM (https://www.cidoc-crm.org/) in combinazione con il modello FRBR, o di RiC-O (https://www.ica.org/standards/RiC/RiC-O_v0-2.html).

[6] I modelli dei casi studio vengono presentati in UML (per ragioni di spazio e visualizzazione, il diagramma qui fornito rappresenta solo parzialmente il caso di studio); i grafi RDF corrispondenti ai diagrammi UML sono disponibili all'indirizzo https://github.com/DFCLAM/bigrafo. Si precisa che allo stato attuale grafi RDF e l'adattamento di FaBiO sono in fase di elaborazione.

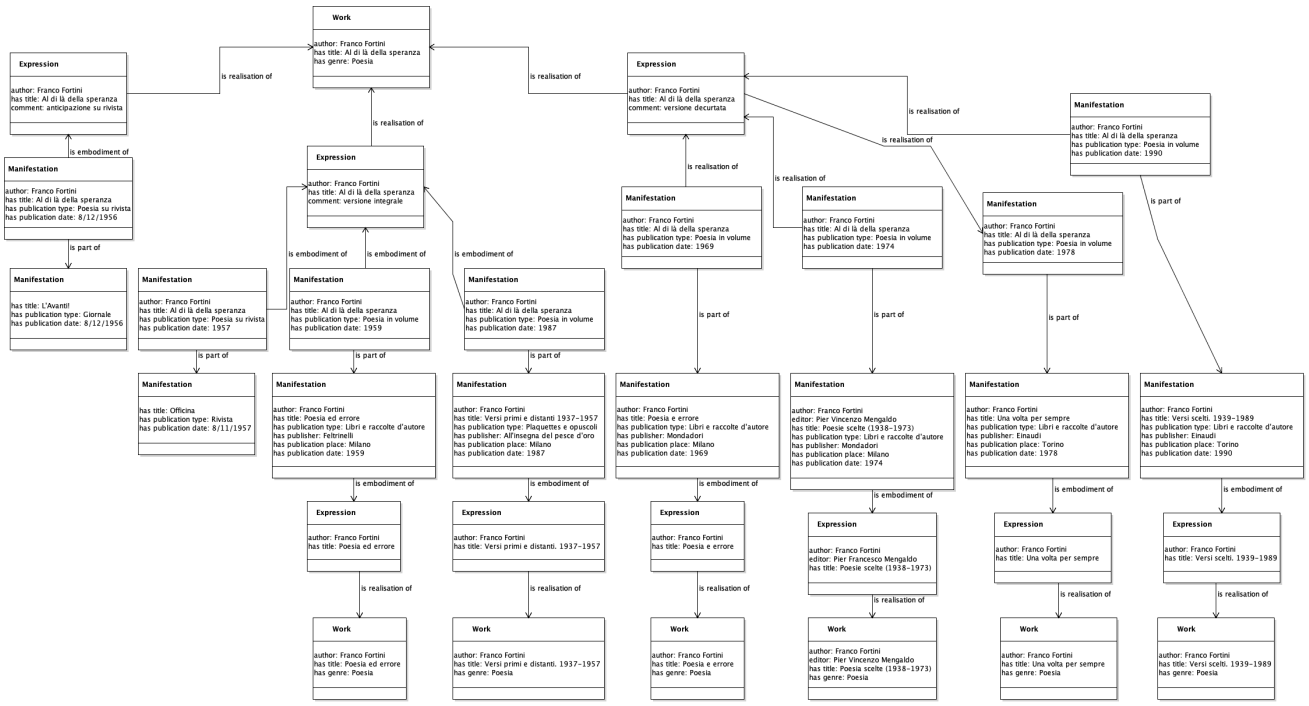


Figura 2. Modello concettuale (parziale) basato su FaBiO – Al di là della speranza

Guardando al diagramma nella figura 2, si nota inoltre la distinzione tra la singola poesia e la raccolta o la rivista dove la poesia è stata pubblicata. Essendo l'ontologia di FaBiO basata su FRBR, abbiamo optato per mantenere questa distinzione ai tre livelli presi in considerazione (*Work, Expression, Manifestation*). Da un punto di vista teorico, questa scelta appare ragionevole poiché non solo per il singolo componimento ma, ad esempio, per un intero volume di poesie possiamo considerare tanto il suo livello documentale (*Work*), quanto la sua componente testuale (*Expression*) ed editoriale *(Manifestation*). D'altra parte, la rappresentazione di proprietà come il titolo, l'autore o il curatore viene ripetuta tra le varie istanze delle entità *WEM* per ragioni pratiche (ad esempio per facilitare la visualizzazione e l'interrogazione dei dati).[7]

Per quel che riguarda le divergenze rispetto a FaBiO, esse riguardando essenzialmente la decisione di non adottare l'ontologia nella sua interezza, date le ambiguità in alcune sue scelte di modellazione. In linea ancora una volta con quanto scrive Biagetti [4, 5], notiamo che in FaBiO sono presentate come sottoclassi di *Work*, ad esempio, *Artistic work, Critical edition, Essay, Report, Research paper, Review*; tra le sottoclassi della classe *Expression*, d'altro canto, troviamo *Article, Book, Chapter, Comment, Letter, Manuscript, Periodical issue, Proceedings paper* e altre: nel contesto di un progetto come BiGraFo non pare del tutto convincente la differenziazione tra un saggio, un'edizione critica e un contributo scientifico come *Work* e un contributo negli atti di un convegno o capitolo di libro come *Expression*.

Le sottoclassi previste da FaBiO per *Work* ed *Expression* non sono inoltre risultate del tutto adeguate a descrivere la produzione di Fortini: FaBiO nasce infatti con il diverso e specifico intento di rappresentare gli aspetti relativi alle pubblicazioni scientifiche, e non la produzione letteraria di autori. Se sottoclassi di *Work* come *Musical composition, Novel, Play, Poem, Screenplay* e *Short story* (a loro volta sottoclassi di *Literary artistic work*) potrebbero con alcuni aggiustamenti adattarsi alla produzione fortiniana (con l'assenza della diaristica e delle parole per musica), risulta più complesso un adattamento di alcuni casi a livello di *Expression*, che non presenta, ad esempio, alcuna sottoclasse idonea a illustrare la relazione esistente tra un componimento poetico e la raccolta in cui è esso contenuto: si potrebbe in alternativa impiegare la sottoclasse *Chapter*, ma questa soluzione non sarebbe esente da forzature. Nel contesto di BiGraFo, quindi, si riutilizza solo una porzione di FaBiO, in particolare la struttura più generale del modello, lasciando aperta la possibilità del suo completo utilizzo qualora questo fosse necessario e qualora le problematicità sopra menzionate venissero risolte. Inoltre, alcune delle informazioni veicolate in FaBiO da queste sottoclassi sono state demandate a nuovi elementi di modellazione introdotti nell'ontologia, ossia la classe *Genre* (importata dall'ontologia di Dbpedia), per indicare il genere (narrativa, poesia, saggistica, pubblicistica, parole per musica, testi per film) a livello di *Work* e la classe *PubblicationType* (libri e

[7] Le relazioni utili a rappresentare l'autore o il curatore sono state importate da risorse esistenti, come schema.org (https://schema.org/).

raccolte d'autore, *plaquettes* e opuscoli, opere in collaborazione, traduzioni in volume, poesie in rivista, poesie in volume, ecc.) per rappresentare le diverse tipologie di pubblicazione a livello di *Manifestation*. Questa strategia ha permesso di articolare e mettere in valore delle informazioni che, anche nella prospettiva dell'interrogazione del modello, risultano semanticamente rilevanti.

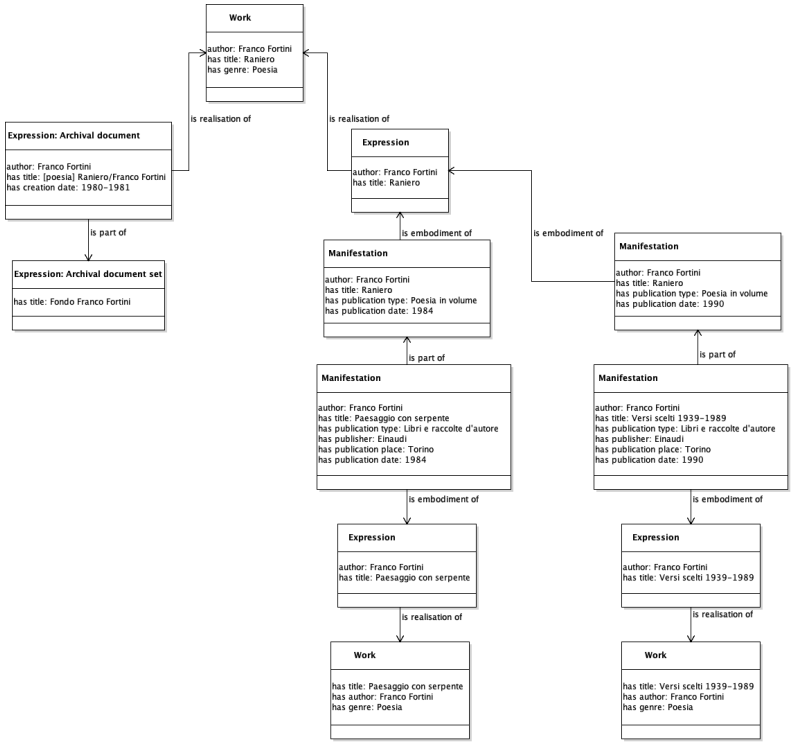


Figura 3. Modello concettuale (parziale) basato su FaBiO – Raniero

Il secondo caso di studio che proponiamo permette di illustrare come sia possibile integrare, attraverso il modello di FaBiO, il dato archivistico: per la poesia *Raniero* (vd. Fig. 3), da un unico *Work* si realizzano due *Expression*: la prima dà luogo, in seguito, a due edizioni (*Manifestation*), rispettivamente nelle raccolte poetiche *Paesaggio con serpente* e *Versi scelti 1939-1989* (Einaudi 1984 e 1990); la seconda veicola il relativo documento d'archivio conservato nel Fondo Franco Fortini. Come si accennava, il progetto non esclude una futura integrazione con un'ontologia archivistica specifica, tuttavia FaBiO consente una prima e essenziale relazione con il patrimonio documentario fornendo alcuni dati utili (*title, creation date e creator* nel caso del documento, *title* nel caso del fondo archivistico).

4. RISULTATI E PROSPETTIVE

A questo stadio ancora iniziale del lavoro, in cui ci si è concentrati sugli aspetti di modellizzazione e sull'analisi delle specificità dei dati, si impone una riflessione sull'impiego di modelli già esistenti e sulla creazione di modelli specifici *ad hoc* per il progetto. In questo caso si è scelto di adottare la struttura principale di un'ontologia, FaBiO, realizzata per un contesto diverso rispetto a quello in cui qui viene applicata, essendo nata per rappresentare gli aspetti relativi alle pubblicazioni scientifiche e non la produzione intellettuale e letteraria d'autore. Se la scelta, da una parte, richiede uno sforzo di adattabilità e in qualche misura un compromesso, dall'altra consente di utilizzare strumenti condivisi, vantaggiosi in termini di sostenibilità e interoperabilità. Anche da un punto vista concettuale si è riflettuto in termini di esportabilità e riutilizzabilità: la nostra riflessione parte dal caso fortiniano (e nel caso più specifico dalla bibliografia d'autore, con la prospettiva di aprirsi ai materiali d'archivio e alla bibliografia della critica), tuttavia prevediamo di verificare l'ipotesi del modello anche nella produzione di altri scrittrici e scrittori del Novecento.

Un obiettivo essenziale sarà pertanto modellizzare un'ontologia che si basi sulla struttura *core* di FaBiO e che al tempo stesso possa rispondere con più precisione alle necessità del progetto (si è visto, come esempio, il caso della rappresentazione del genere e delle tipologie di pubblicazione).

Teniamo inoltre a sottolineare che le considerazioni e le riflessioni qui presentate hanno, per il momento, un carattere preliminare ed esplorativo: l'ontologia è in fase di elaborazione, in confronto costante con gli studiosi di Fortini e con gli esperti di ontologia applicata, nella convinzione che ogni buona operazione di modellizzazione costituisca un processo dinamico e iterativo, che evolve a seconda delle esigenze del progetto e migliora all'interno di un dialogo costante.

5. RINGRAZIAMENTI

Il progetto *Biblio-grafo. Un catalogo semantico per il Centro di ricerca Franco Fortini* (BiGraFo) è finanziato dal Piano di Sostegno alla Ricerca 2022 dell'Università di Siena per progetti "Curiosity-driven" (F-CUR). Il gruppo del progetto include Niccolò Scaffai (direttore del Centro Interdipartimentale di ricerca Franco Fortini), Stefano Moscadelli, Riccardo Castellana e Marco Maggini. Si ringrazia per il contributo essenziale e per le riflessioni condivise Eleonora Bassi, direttrice della Biblioteca di Area Umanistica dell'Ateneo, l'archivista Elisabetta Nencini, e Luca Lenzini, coordinatore del Centro Fortini.

BIBLIOGRAFIA

[1] Baker, Thomas, Karen Coyle, e Sean Petiya. «Multi-entity models of resource description in the semantic web: A comparison of FRBR, RDA and BIBFRAME». *Library Hi Tech* 32, fasc. 4 (2014): 562–582.

[2] Bassi, Eleonora, e Elisabetta Nencini, (a cura di). *Bibliografia di Franco Fortini*. Macerata: Quodlibet, 2022.

[3] Bekiari, Chryssoula, Martin Doerr, Patrick Le Bœuf, e Pat Riva. *LRMOO object-oriented definition and mapping from IFLA LRM*. Version 0.9.6., 2023. https://cidoc-crm.org/frbroo/sites/default/files/LRMoo_V0.9.6.pdf

[4] Biagetti, Maria Teresa. «A Comparative analysis and evaluation of bibliographic ontologies». In Challenges and Opportunities for Knowledge Organization. The Digital Age. Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal, a cura di Fernanda Ribeiro e Maria E. Cerveira, 501–510. Ergon: Baden Baden, 2018.

[5] Biagetti, Maria Teresa, (a cura di). *Le ontologie bibliografiche: modelli concettuali e vocabolari condivisi per l'universo bibliografico*. Roma: Bulzoni, 2022.

[6] Coyle, Karen. «Works, expressions, manifestations, items: an ontology». *Code4Lib Journal,* 53 (2022). https://journal.code4lib.org/articles/16491

[7] De Berardinis, Jacopo, Valentina Anita Carriero, Albert Meroño-Peñuela, Andrea Poltronieri, e Valentina Presutti. «The Music Meta Ontology: a flexible semantic model for the interoperability of music metadata». In *Proc. of the 24th Int. Society for Music Information Retrieval Conf.*, 859–867. Milano, 2023.

[8] Holden, Chris. «The Bibliographic Work: History, Theory, and Practice». *Cataloging & Classification Quarterly* 59, fasc. 2–3 (2020): 77–96. https://doi.org/10.1080/01639374.2020.1850589.

[9] Jett, Jacob, Terhi Nurmikko-Fuller, Timothy Cole, Kevin Page, e J. Stephen Downie. «Enhancing Scholarly Use of Digital Libraries: A Comparative Survey and Review of Bibliographic Metadata Ontologies». In *IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, a cura di Nabil R. Adam, Lillian (Boots) Cassel, Yelena Yesha, Richard Furuta, e Michele C. Weigle, 35–44. Newark, NJ, USA, 2016.

[10] Lisena, Pasquale, e Raphaël Troncy. «Representing complex knowledge for exploration and recommendation: the case of classical music information». *Applications and Practices in Ontology Design, Extraction, and Reasoning* 49 (2020).

[11] Lucarelli, Anna, Eleonora Marzocca, Elisabetta Viti, Pino Buizza, Alberto Cheti, Luciana Franci, Maria Chiara Giunti, e Marta Ricci, (a cura di). *Biblioteca nazionale centrale di Firenze. Nuovo soggettario. Guida al sistema italiano di indicizzazione per soggetto. Seconda edizione interamente rivista e aggiornata*. Roma, Firenze: Associazione italiana biblioteche, Biblioteca Nazionale Centrale di Firenze, 2021. https://doi.org/10.53263/9788878123465

[12] Mandal, Sukumar. «Item Relationships using Dublin Core, BIBO, FOAF, and FRBR for Managing Resources of Cultural Heritage: Designing a Prototype Integrated Framework». *Library Philosophy and Practice (e-journal)*, 2021. https://digitalcommons.unl.edu/libphilprac/5329

[13] Meschini, Federico. *Oltre il libro. Forme di testualità e Digital Humanities*. Milano: Editrice Bibliografica, 2020.

[14] Peroni, Silvio, e David Shotton. «FaBiO and CiTO: Ontologies for Describing Bibliographic Resources and Citations». *Web Semantics: Science, Services and Agents on the World Wide Web* 17 (2012): 33–43. https://doi.org/10.1016/j.websem.2012.08.001

[15] Peroni, Silvio, e David Shotton. «The SPAR Ontologies». In *The Semantic Web – ISWC 2018.*, a cura di Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, e Elena Simperl, 11137:119–36. Lecture Notes in Computer Science. Springer International Publishing, 2018. https://doi.org/10.1007/978-3-030-00668-6_8

[16] Peroni, Silvio, David Shotton, e Fabio Vitali. «Scholarly Publishing and Linked Data: Describing Roles, Statuses, Temporal and Contextual Extents». In *Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12)*, 9–16. New York: ACM Press, 2012. https://doi.org/10.1145/2362499.2362502.

[17] Sanfilippo, Emilio M. «Ontologies for information entities: State of the art and open challenges». *Applied Ontology* 16, fasc. 2 (2021): 111–135.

[18] Smiraglia, Richard P. «Musical Works as Information Retrieval Entities: Epistemological Perspectives». In *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR 2001)*, 85–91. Bloomington, 2001. https://doi.org/10.5281/zenodo.1416512.

[19] Tomasi, Francesca. *Organizzare la conoscenza: Digital Humanities e Web Semantico*. Milano: Editrice Bibliografica, 2022.

Lost in datification? The journey of data from the primary source to the final interpretation

Enrica Bruno[1], Sofia Baroncini[2], Francesca Tomasi[3]
[1] University of Bologna, Italy - enrica.bruno2@unibo.it
[2] Leibniz Institute for European History (IEG), DH Lab, Germany - baroncini@ieg-mainz.de
[3] University of Bologna, Italy - francesca.tomasi@unibo.it

ABSTRACT[1]

In the field of Digital Humanities, recent attention was given to the relationship between RDF triples and natural language in the context of natural language to RDF conversion of humanities texts. The rigid structure of ontologies obliges scholars to make critical choices during the formalization of data resulting from an interpretation of the cultural resource. This may result in crucial differences between the final RDF formalization and the natural language text in terms of how much of the final semantic content retains the original one and how much remains hidden due to the framework's formal structure. The verification of the adherence of structured data to the primary source is useful to test if the data model returns the semantic expressiveness of the primary source in order to pursue the specific goal driven by the computational process. In the context of humanities, in which the precision of information is fundamental for addressing sound analyses, this verification becomes crucial when derived data are treated as a substitute for the primary source in computational tasks. In this talk, we propose a three-step approach to verify the extent to which the RDF triples represent the respective content of the textual source from which they were generated within the limits of the modeling adopted. The approach is thus tested by proposing two case studies taken from two different cultural domains, namely literature and art history.

KEYWORDS

RDF data for humanities; RDF data quality; literature; art history.

1. INTRODUCTION

In the digital humanities field, growing attention is given to the issue of expressing the semantic content of texts (i.e., primary sources or their scholarly interpretation) in structured, computer-processable formats. Such formalization implies a selection of the information carried by the textual sources that will be expressed according to the specific perspective adopted during the modeling [19]. Nevertheless, even considering the adopted scope only, the process of transformation into structured data may lead to a manipulation of the primary information at various levels. Not only the actual content expressed may be misinterpreted, but also the ontological modeling adopted may affect its semantics. Furthermore, a user should contextualize data (i.e., how it was created and from which source) to avoid misleading interpretations during the data analysis. Assuming that structured data may be used as a faithful representation of the primary source in several computational analyses and information retrieval systems, the definition of the degree of difference between the resulting data and the source gains crucial importance. Despite several methodologies to verify ontology and triple readability being available [6, 18], to the extent of the author's knowledge, none of them addresses the topic of semantic alteration in the overall process of manually generating data from humanities texts. We state that this domain deserves particular attention to the accuracy of the created data for two reasons. First, due to the primary source complexity, the manual data creation does not guarantee its complete accuracy, as is generally accepted in the computer science field, since the transformation into data often requires a high degree of the annotator's interpretation, making the conversion a subjective, challenging task. Secondly, the consultation of the primary source in humanities practice has a crucial role in guaranteeing the trustworthiness of the analysis the humanist conducts. For this reason, in the context of the translation of the humanist's task to a computational one, the data should guarantee a level of accuracy equal to the primary sources from which it was extracted. Furthermore, this verification becomes crucial when derived data are treated as a substitute for the primary source in computational tasks.

In this talk, we propose a three-step approach to verify the extent to which the RDF triples represent the content expressed by the textual source from which they were generated within the limits of the modeling perspective adopted.

[1] Sofia Baroncini is responsible for Introduction and Conclusion. Enrica Bruno and Sofia Baroncini are responsible for the State of Arts, Methodology and Case Studies sections. Francesca Tomasi was the scientific supervisor providing critical revisions and feedback for the research and the entire writing process.

The main Research Question (RQ) can be expressed as follows: *to what extent do the RDF data and the resulting representations preserve the content of the textual source within the limits of the perspective adopted by the modeling?*

To answer this question, we identify three levels at which a semantic alteration may occur in the process of conversion of a humanist text into RDF data:

Meaning: is the semantic meaning expressed by the structured data adherent to the source (RQ1)?
Semantic modeling: are the modeling choices semantically expressive (RQ2)?
Contextualization: is the final RDF data contextualized, viz., is the reference to the provenance of data and statement in general indicated (RQ3)?

For each level, we provide an approach for evaluation adapted from metrics available in the state of the art, and we test it over two case studies from two diverse scenarios: the RDF conversion of 1) Italo Calvino's collection *Il castello dei destini incrociati* and 2) some of the art critique texts by the art historian Erwin Panofsky. The case studies were chosen as they belong to two different domains, namely literature and art history, and two respective types of sources (i.e. direct source and an interpretation of the primary source) to guarantee a wide range of applications in the broad humanities field.

2. STATE OF ART

Evaluation of quality is a crucial aspect when dealing with data. It aims at guaranteeing the correctness of data from multiple points of view, ranging from the logical aspects to the accuracy of the content expressed. The ISO standard related to data quality[2] defines it as the "degree to which data satisfy the requirements defined by the product-owner organization" in relation to 15 characteristics of the data. Furthermore, W3C provides a framework for data quality description[3], and multiple approaches for assessing semantic data quality data are present in the literature. Among them, worth mentioning [1], who extends data quality to every stage of the creation process, and [5], who provides a definition for 44 measures on the basis of the ISO definitions tested over large RDF datasets. Although many dimensions can be evaluated with an automatic or semi-automatic approach, some of them require human validation or the aid of domain experts, especially for provenance and contextual information criteria related to the specific use case [8, 4].

Several approaches providing ontology quality evaluation exist [6, 18] together with validator tools to ensure that ontologies are well-formed, consistent, and adhere to established methods (e.g. OWL API[4] Validator, Protégé Ontology Validator[5], and RDF Validator[6]). The validation of ontology semantics is usually done instead by involving domain experts as validators [17].

Various tools for converting text to RDF or the contrary are then currently available, showing an increasing accuracy in their results. Nevertheless, to the authors´ knowledge, they are not feasible for the task of validating the conversion of scholarly and literary text into RDF data. The quality of the conversion of text to RDF may be highly influenced by the writing style adopted and the absence of a specific structure, such as a database format [7, 16].

Furthermore, traditional RDF-to-text methods are based on hand-crafted and domain-specific rules of conversion. These methods provide a readability of the structured data [20, 11] where users are considered as testers and evaluators of the data model in order to verify an alignment between the user representation and the formal annotation [2], but it doesn't ensure the validity of the triple content. Despite algorithms based on showing promising results being available, their evaluation of correctness does not reach a sufficient score for the scope of this article [20]. For this reason, we adopt a human-based verification approach.

3. METHODOLOGY

As the verification occurs on the content level, the validators are experts in the domain of Digital Humanities, specifically semantic web technologies. The understandability of RDF data from generic users is not in the scope of this paper. The validation was performed by one validator for each case.

First step(RQ1): According to [1: 5] the semantic validity of triples is proved whether (i) it is available from a trusted source, (ii) it is common sense or (iii) the stated property can be directly measured. We adapt the metric by evaluating

[2] https://iso25000.com/index.php/en/iso-25000-standards/iso-25012
[3] https://www.w3.org/TR/vocab-dqv/
[4] https://owlapi.sourceforge.net/
[5] https://protegewiki.stanford.edu/wiki/Validation
[6] https://www.w3.org/RDF/Validator/

whether each triple from the case study expresses content that holds true when compared with the source text. The triple scores 1 if its content is fully represented, 0,5 if it is partially represented or inaccurate, and 0 if it is absent.

Second step(RQ2): Although several methodologies for ontological semantic validity based on experts' evaluation exist [6],

our focus is to understand whether the adopted ontology is suitable to express the semantic content of the specific case study. To this end, we evaluate whether the type of entity assigned to each instance is suitable for describing the content of the phrase to which they refer. Therefore, all the triples having a relation for declaring the type of entity are extracted, and the accuracy of the type assigned is evaluated against the content provided by the text. The metric scores 1 if the class fully represents the content corresponding to the subject of the triple, 0,5 if it partially expresses it, and 0 if it does not express it.

Third step(RQ3): Contextual references concerning RDF triples are necessary to enable the verification of data [17], the assessment of reliability [9], and the analysis of the processes that generated the data [10]. For this respect, RDF data contextualization verifies whether the following aspects are made explicit through one or multiple statements:

The entity responsible for the intellectual content of the resource (i.e., the creator)
The indication of the source from which data were extracted
Contextual information about the statement provenance

In this study, the evaluation is carried out on three possible levels (i.e., triple, entity, and graph), as our aim is to evaluate only whether the information is present. Similarly, the presence of contextual information expressed with different strategies (e.g. if the creator attribution is present both at the graph and statement level) does not affect the final score. Nevertheless, if the contextual information is provided at the assertion level, it should be verified whether every assertion examined has such information.

The first point scores 1 if the creator is stated and 0 if it is not. As regards the second point, the metric considers degrees of details, adding to the total score a) 0,5 if only the resource is cited, b) 0,25 if there is more specific information about the portion of text from which the triple is generated, and c) 0,25 if the text reference is present. Concerning the third point, the metric scores 0,75 if the responsibility of the statement is declared, as it is the fundamental aspect of provenance information. To this score, 0,25 is added if there is more information detail (e.g., the time when the statement was created).

4. CASE STUDIES

Here, we propose two case studies to test the three-step approach proposed to verify the extent to which the RDF triples properly represent the textual source from which they were generated. Both of the case studies are taken from two different cultural domains, namely literature and art history, and are resources manually created by domain experts. The main differences can be summarized in the table below (see Tab. 1).

BACODI	Iconology dataset
Literature	Art history
Direct source	Interpretation of the primary source
Description of the narrative-combinatorial relations between tarot cards in the collection's stories	Report of the work's interpretation (iconography and meanings)
Ontology *ad hoc*	Ontology based on theoretical approach, reused ontologies
Text with a defined structure	Discursive text

Table 1. Summary of the characteristics of the two case studies

BACODI (Base di Conoscenza dell'Ontologia dei *Destini incrociati* di Italo Calvino)

ODI (Ontologia dei Destini incrociati di Italo Calvino) and its corresponding Knowledge Base BACODI (Base di Conoscenza dell'Ontologia dei Destini incrociati di Italo Calvino) were created to represent and describe the first edition of Italo Calvino's *Il castello dei destini incrociati* [3] together with the description of the tarot cards used by the author to create twelve stories in the first homonymous collection of the work. In particular, BACODI stored the description of tarot cards considered both as cultural artifacts and as narrative instances in the text. The double descriptive dimension of each

tarot card combines the literary nature of the textual resource with its particular structure adopted by the author, who uses cultural objects from the artistic domain for storytelling[7].
The case study reported here is the fourth story of the collection, *Storia d'un ladro di sepolcri*, and in particular, the description of some tarot cards as narrative instances in the text, thus also considering their relationships within the story.

The Iconology dataset
The Iconology (ICON) dataset[8] represents the art interpretations by the art historian Erwin Panofsky expressed in four of his major contributions [12, 13, 14, 15]. For this purpose, an ontology based on his own theory was created (ICON ontology[9]). The interpretations concern ca. 400 artworks, mainly from the Western Early Modern period. Interpretations identify subjects and meanings depicted in the artworks according to the art historian and are distinguished on three levels of understanding, from a more superficial to a deeper one. In this way, objects, iconographies, and meanings are described, along with further details, if any (e.g., the textual or visual evidence supporting the recognition), and the provenance of the assertion.
The case study selected is the interpretation of a relief on the external wall of Modena Cathedral, which represents either Cupid or a personification of Death. The case is discussed in a chapter in which the tendency of the Middle Ages to read classical figures with a moral implication is treated.

5. RESULTS[10], CONCLUSION, AND FUTURE WORK

Table 2 shows the results of the analysis. Both datasets performed high scores in the first two steps of the evaluation, showing lower results only in the contextualization part (scores: 0,67 and 0,83). This is due to the fact that, on one hand, BACODI does not provide provenance information despite providing very precise indications of the creator of the content and of the text portion from which the information was extracted. On the other hand, the Iconology dataset provides provenance information, but it is not precise when indicating the text source, as only the reference to the overall book is provided.
Furthermore, despite showing similar results for the meaning aspect (0,95, and 0,96), details noted by the validators during the analysis need further consideration. In a few cases, the meanings noted in the triples were the result of a human interpretation of implicit knowledge. For instance, it is stated that the protagonist of Calvino´s story desires richness. Although this fact is never explicitly asserted by the text, it can be understood by reading the overall story. In other cases, the meanings are forged to facilitate modeling. For instance, in the Iconology dataset, it is stated that the artifact represents the action *grabbing with one hand,* whereas in the text, the phrasal verb *carrying* is used instead of *grabbing with*. This choice can be justified as a need to assimilate terms with more than one occurrence. As regards modeling choices, BACODI's results show an appropriate selection of the types for each instance, whereas the iconology dataset has little differentiation between entities represented in the artifact chosen (i.e. "torch", "putto", "ibis" and "wreath" are all in the class `icon:NaturalElement`). This is motivated by the different modeling scopes of the two cases. Whereas ODI was created to model Calvino´s specific book, the ICON ontology formalizes the domain of iconographical and iconological interpretations, aiming at being suitable for further domain descriptions.

	Results of step 1 (Meaning)	Results of step 2 (Semantic modeling)	Results of step 3 (Contextualization)
BACODI	0,95	1	0,67
Iconology dataset	0,96	0,85	0,83

Table 2. Results of the evaluation

[7] The complete documentation can be found at https://odi-documentation.github.io/materials/
[8] Available at https://iconology-dataset.streamlit.app/
[9] Available at https://w3id.org/icon/docs
[10] A complete overview of the results is available at https://doi.org/10.5281/zenodo.10973092

This talk focused on the semantic expressiveness of RDF triples, trying to question the extent to which data formalized represents the respective content of the native textual source. As the humanities texts are characterized by aspects of complexity (e.g., implicit, articulated, or undefined knowledge), their manual translation into data may be subject to variations at multiple levels. This aspect is crucial when considering the importance that the consultation of primary sources has for the humanities research advancement. The goal was to provide an approach to assess the extent to which semantic information manually extracted from humanities texts is valid, in relation to the layers of content, semantics, and contextualization. Being evaluated over case studies from two different domains and with different characteristics, we argue it can be applied to text belonging to different domains, either on a primary source or on scholarly literature. The fact that it is based on human evaluation allows us to verify also the modeling of implicit knowledge, which is likely to be embedded in complex texts such as literary and scholarly ones.

The current evaluation is limited to the extent to which the selected information provided by the text is correctly retained, expressed, and contextualized in data. Nevertheless, the semantic modeling of a resource aims not only at mirroring the plain text's content but also at enriching it with further knowledge that may be seen by a domain expert [19]. Moreover, a more extensive measurement involving multiple evaluators is needed to reduce the impact of evaluators' subjectivity on the results. Although the proposed approach does not aim to be an exhaustive solution to the challenging problem of derived data accuracy, it faces an initial critical reflection on the topic, proposing a practical method that can be applied in DH projects involving complex humanistic textual sources. Future work includes: 1) a more extensive evaluation including multiple evaluators and further domains, and 2) the extension of the approach to measure the added semantic enrichment that an expert of the considered domain may have included in the dataset during its creation process, as it constitutes a core aspect of the digital humanists' practice.

REFERENCES

[1] Assaf, Ahmad, and Aline Senart. 'Data Quality Principles in the Semantic Web'. In *2012 IEEE Sixth International Conference on Semantic Computing*, 226–29, 2012. https://doi.org/10.1109/ICSC.2012.39

[2] Bonora, Paolo, Martina Dello Buono, Francesca Giovannetti, and Francesca Tomasi. 'Tell Me the Truth. Validating the Semantic Alignment between the Annotation User Interface and the Knowledge Base'. In *Digital Humanities 2023: Book of Abstracts*, 202–4. Zentrum für Informationsmodellierung-Austrian Centre for Digital Humanities, University of Graz, 2023. https://doi.org/10.5281/zenodo.7961822.

[3] Calvino, Italo. *Il castello dei destini incrociati*. Torino: Einaudi, 1973.

[4] Daquino, Marilena, Valentina Pasqual, and Francesca Tomasi. 'Knowledge Representation of Digital Hermeneutics of Archival and Literary Sources'. *JLIS: Italian Journal of Library, Archives and Information Science = Rivista Italiana Di Biblioteconomia, Archivistica e Scienza Dell'informazione: 11, 3, 2020*, no. 3 (2020): 59–76. https://doi.org/10.4403/jlis.it-12642.

[5] Färber, Michael, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. 'Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO'. Edited by Amrapali Zaveri, Dimitris Kontokostas, Sebastian Hellmann, and Jürgen Umbrich. *Semantic Web* 9, no. 1 (2017): 77–129. https://doi.org/10.3233/SW-170275

[6] Gangemi, Aldo, Carola Catenacci, Massimiliano Ciaramita, and Jos Lehmann. 'Modelling Ontology Evaluation and Validation'. In *European Semantic Web Conference*, 140–54. Berlin: Heidelberg: Springer Berlin Heidelberg, 2006. https://doi.org/10.1007/11762256_13

[7] Hassanzadeh, Kimia, Marek Reformat, Witold Pedrycz, Iqbal Jamal, and John Berezowski. 'T2R: System for Converting Textual Documents into RDF Triples'. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 3:221–228, 2023. https://doi.org/10.1109/WI-IAT.2013.187

[8] Iyer, Vivek, Lalit Mohan Sanagavarapu, and Y. Raghu Reddy. 'A Framework for Syntactic and Semantic Quality Evaluation of Ontologies'. In *Secure Knowledge Management In The Artificial Intelligence Era*, edited by Krishnan Ram, H. Raghav Rao, Sanjay K. Sahay, Samtani Sagar, and Zhao Ziming, 73–93. Communications in Computer and Information Science. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-97532-6_5

[9] McGlothlin JP, Khan L. 'Efficient RDF Data Management Including Provenance and Uncertainty'. In *Proceedings of the Fourteenth International Database Engineering and Applications Symposium*, 193–198. ACM, New York, 2010. https://doi.org/10.1145/1866480.1866508

[10] Moreau, L. 'The Foundations for Provenance on the Web'. *J Found Trends Web Sci* 2, no. 2–3 (2010): 99–241. https://doi.org/10.1561/1800000010

[11] Moussallem, Diego. 'Knowledge Graphs for Multilingual Language Translation and Generation'. *ArXiv Preprint ArXiv:2009.07715*, 2020.

[12] Panofsky, Erwin. *Meaning in the Visual Arts*. Garden City, NY: Doubleday, 1955.

[13] Panofsky, Erwin. *Renaissance and Renascences in Western Art*. New York: Harper & Row, 1972.

[14] Panofsky, Erwin. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance.* Boulder, Colo: Westview Press, 1972.

[15] Panofsky, Erwin, and Fritz Saxl. 'Classical Mythology in Mediaeval Art'. *Metropolitan Museum Studies* 4, no. 2 (1933): 228–280. https://doi.org/10.2307/1522803

[16] Rincon-Yanez, Diego, and Sabrina Senatore. 'FAIR Knowledge Graph Construction from Text, an Approach Applied to Fictional Novels'. In *Proceedings of the 1st International Workshop on Knowledge Graph Generation from Text and the 1st International Workshop on Modular Knowledge Co-Located with 19th Extended Semantic Web Conference (ESWC 2022)*. Hersonissos, Greece, 2022.

[17] Sikos, L.F., and D. Philp. 'Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs'. *Data Sci. Eng.* 5 (2020): 293–316. https://doi.org/10.1007/s41019-020-00118-0

[18] Syed, Zafar Habeeb, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 'Factcheck: Validating Rdf Triples Using Textual Evidence'. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018*, 2018. https://doi.org/10.1145/3269206.3269308

[19] Tomasi, Francesca. *Organizzare La Conoscenza: Digital Humanities e Web Semantico*. IT: Editrice Bibliografica, 2022. https://doi.org/10.53134/9788893573573

[20] Zhu, Yaoming, Juncheng Wan, Zhou Zhiming, Liheng Chen, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 'Triple-to-Text: Converting RDF Triples into High-Quality Natural Languages via Optimizing an Inverse KL Divergence'. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 455–464, 2019. https://doi.org/10.1145/3331184.3331232.

Orbis dioecesium. Creating authority data on the legal-historical changes of Catholic dioceses

Benedetta Albani[1], Rowan Dorin[2], Yohan Park[3]
[1]Max Planck Institute for Legal History and Legal Theory, Germany - albani@lhlt.mpg.de
[2]Stanford University, United States - dorin@stanford.edu
[3]Max Planck Institute for Legal History and Legal Theory, Germany - park@lhlt.mpg.de

ABSTRACT

In this paper, we address the publishing of authority data on the legal-historical evolution of Catholic dioceses fulfilling FAIR-Principles. The paper discusses the methodology of building the domain-specific ontology model and publishing authority data using the knowledge base Wikibase. The paper is divided into four parts. The first part outlines the aims and goals of the project and the state of the art on semantic modelling of historical data and its implementation as knowledge graph (KG). The second part presents each phase of the project: the elaboration of the data set, the development of an event-based ontology model, and the principles that guided the creation and maintaining of the items and properties in the Wikibase environment. The fourth part describes the pipeline process for data ingestion. The conclusion presents possible scholarly and general uses of the project.

KEYWORDS

Modelling Historical Changes; Catholic Dioceses; Linked Open Data; Knowledge Graph; Wikibase.

1. WHY DO WE NEED AUTHORITY DATA ON CATHOLIC DIOCESES? *STATUS QUAESTIONIS* BETWEEN HISTORICAL DISCIPLINES AND DIGITAL HUMANITIES

Since late antiquity, historical sources have used the names of Christian dioceses to define and identify various elements such as institutions, individuals, locations, documents, and actions. In the field of archives, diocesan names have frequently served as organizing criteria, profoundly influencing information accessibility, transmission and processing, as well as knowledge production. Although the significance of the diocesan system may not be readily apparent in the daily lives of most people today, it holds fundamental relevance for those studying the history of the Christian world and its interactions with different cultures and societies. In fact, a diocese encompasses a complex and multifaceted entity, extending well beyond the mere definition of its territorial extension. It signifies a unique relationship with the Apostolic See and secular authorities, the presence of a specific hierarchy, a particular community, distinct religious practices, and various jurisdictions; it furthermore has implications for the juridical, spiritual, and economic spheres. These elements evolve over time, subjecting each diocese to significant legal-historical changes and making Catholic dioceses an important diachronic actor, from antiquity to the present day.

The extensive literature on the history of dioceses is well-acknowledged, often driven by institutional needs and a focus on episcopal chronotaxis[1]. Specialist works frequently adopt a local perspective, however, concentrating on a specific diocese or those within a particular geographical area. Information on dioceses and their bishops is also continually updated and disseminated by governing institutions, and such data are also accessible online through various media, both institutional and non-institutional, serving diverse purposes such as scientific research, dissemination, and administrative functions. These data present several problems, both from a scientific and technical point of view. First, since they have been collected with different methodologies, they show very heterogeneous levels of accuracy and analytical rigour and are therefore not suitable for comparisons or interoperability. Second, they often fall short of an adequate set of authoritative scientific references (sources, academic literature). Third, they lack adequate forms of standardization, both as regards the naming of dioceses and the description of their various legal-historical changes. Finally, the legal-historical evolution of the dioceses is always described only in a discursive manner, thus limiting unambiguous identification of entities and impeding the systematic description of the various legal-historical changes and the resulting different historical phases of diocesan evolution.

[1] The examples are numerous. We mention here only a few representative texts of a historiographical tradition: Echeverria, Lamberto de [10]; Oviedo Cavada, Carlos and Marciano Barrios Valdés [16]; Jarry, Eugène and Jean-Rémy Palanque [13]; Plongeron, Bernard and Jean-Rémy Palanque [17]; Bravo Ugarte, José [4]; De Sandre Gasparini, Giuseppina et al. [7]; [12]. For a partial historiographical overview of the Italian area see Battelli, Giuseppe [2: 391-426].

To address these problems, our project aims to elaborate authority data with controlled vocabularies and a taxonomy, enabling the identification of dioceses and tracking their legal-historical changes from the date of the erection of each diocese to the present. The project employs an ontology model and linked data format, ensuring interoperability and adherence to Linked Open Data principles, in a way similar to Wikidata. From the perspective of Digital Humanities, our project aims to answer the following scientific questions:

- How can we model historical changes, namely representing the legal-historical evolution of Catholic dioceses in a standardised model?
- From the perspective of Linked Open Data (LOD), what technical or software solutions can be used to provide users with authority data that meets the FAIR guidelines?

As for the first point, there are many works on theoretical and practical approaches to semantic modelling of historical spatiotemporal information. Since our platform uses an event-based ontology model [14], we consulted several examples during its design, including Simple Event Model (SEM) [11], the Italian Cultural Heritage knowledge graph, ArCo [5,6] as well as CIDOC CRM [9], which is widely used as upper-model in the modelling of semantic historical data. Lastly, we also took into account the adaptation of the Semantic Data for Humanities and Social Sciences CIDOC CRM Extension (SDHSS) [3]. Regarding the second point, it is well known that the vast amount of information on the World Wide Web has paradoxically led to the difficulty of retrieving domain-specific information. As an antidote, knowledge graphs (KG) have arisen to serve as a shared substrate for constantly evolving domain knowledge within a particular community [8]. Recently, there has been a lot of research on the implementation of knowledge bases and knowledge graphs to access information about domain-specific data [15]. The most well-known software infrastructure for the implementation of knowledge graphs is the Wikibase environment provided by the Wikimedia Foundation, a set of software behind Wikidata. This environment allows the implementation of graph data models for structured data, in particular internally relational data models and interfaces.[2] Recent projects show that the Wikibase ecosystem fulfills the FAIR principles when the implementation of knowledge graphs implies the publishing of information as linked open data.[3]

2. COLLECTING AND MODELLING DATA ON LEGAL-HISTORICAL CHANGES IN CATHOLIC DIOCESES

The project is structured in three phases: 1) Dataset elaboration; 2) Data modelling and ontology development; 3) Creation and maintenance of the items and properties in the Wikibase environment.

A) ELABORATION OF THE DATASET

The dataset underlying the project was compiled by combining tabular datasets created for two different research projects[4] and personally elaborated by members of the research groups on the basis of archival documents, specialised academic literature and scientific research.

The dataset is composed of 1120 Catholic diocesan-like institutions including 852 dioceses, 210 metropolitan archdioceses, 63 archdioceses, 6 patriarchates, as well as 221 ecclesiastical provinces. To describe the historical-legal evolution of the dioceses already included in the dataset, 6012 legal changes have been used so far. For the period up to about 1750, the database is mostly complete for Europe and America, while it is still incomplete for the other continents. Considering the

[2] Fauconnier, Sandra, Dragan Espenschied, Lyndsey Moulds, and Lozana Rossenova. "Many Faces of Wikibase: Rhizome's Archive of Born-Digital Art and Digital Preservation – Wikimedia Foundation". Wikimedia Blog (blog), 2018. https://wikimediafoundation.org/news/2018/09/06/rhizome-wikibase/.

[3] There are already many examples of Digital History projects using Wikibase as a tool for publishing knowledge graphs. *FactGrid. A database for historians* (https://database.factgrid.de/wiki/Hauptseite) hosted by the University of Erfurt, Germany, and coordinated by the Gotha Research Centre, started with 16,000 data records from the Gotha Illuminati Study in 2018 and had more than half a million entries as of 11 May 2023. *Enslaved. People of the Historical Slave Trade* (https://enslaved.org/) [19] is a Wikibase-based research platform that provides a one-stop shop for integrated and structured historical data on the transatlantic slave trade after the Early Modern Period. *MiMoTextBase* (https://data.mimotext.uni-trier.de/wiki/Main_Page) [18], a history of literature domain knowledge graph that forms part of the Text Mining and Modelling (2019-2023) project, makes available bibliographic data on approximately 2000 18th-century French Enlightenment novels extracted with text mining techniques, using a Wikibase environment and queryable via a SPARQL endpoint.

[4] The first project, *Digital Atlas of Dioceses and Ecclesiastical Provinces in Late Medieval Europe (1200-1500)* (https://corpus-synodalium.com/digital_atlas/), directed by Rowan Dorin, was created in tandem with *Corpus Synodalium* (https://corpus-synodalium.com/), a searchable online database of local ecclesiastical legislation in medieval Europe. The second project, *SCC Explorer. An interactive Platform on the History of the Congregation of the Council* (https://www.lhlt.mpg.de/3474803/rg-albani-SCC-explorer/), directed by Benedetta Albani, focuses on the activity of one of the most important dicasteries of the Roman Curia, the Congregation of the Council, which had jurisdiction over the entire Catholic world in modern and contemporary times. Both projects have over time developed datasets on Catholic dioceses in different historical periods.

suppression of old dioceses, the unification, and the erection of new ones, we estimate that our dataset comprises approximately 33% of the Catholic dioceses existing today.[5] On this basis, we will continue to enrich the dataset through collaborative methods with specialists to improve its completeness and authoritativeness.
The dataset is currently recorded as CSV tabular data and each legal-historical change is sorted according to the name of the corresponding diocese in alphabetical order. Additionally, beyond bibliographical references, it includes primary sources like papal and episcopal decisions, decrees and concordats between secular states and the Holy See that directly produced legal changes, such as the union of dioceses and changes in jurisdiction. As of January 2024, we have ingested 745 Catholic dioceses from its premise dataset into the platform, and the country distribution[6] of Catholic dioceses in the current dataset is as follows (see Fig. 1).

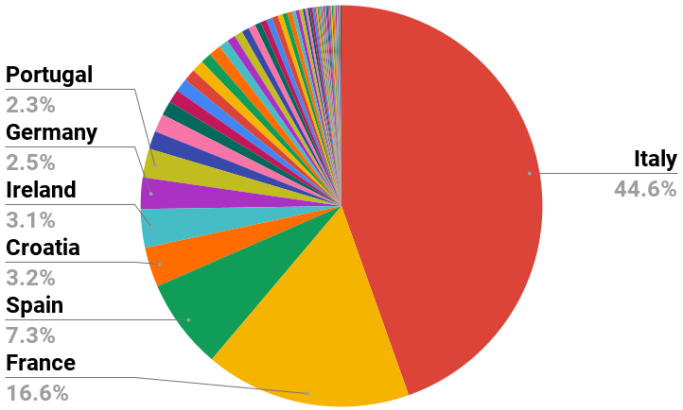


Figure 1. Distribution of modern countries currently included in the dataset, according to current/past location of the diocesan seat

In view of the possibility of participating in the AIUCD 2024 - *MeTe digitali* conference, we are modelling, analysing data and providing visualisations concerning dioceses in the Mediterranean area. As it is well known, the Mediterranean is a crossroad of diverse civilizations, facilitating cultural exchanges, trade, and interactions. In particular, the religious diversity of the Mediterranean exercised significant influence on the historical evolution of dioceses, which can therefore serve as frameworks for studying the coexistence and competition between different cultural traditions.

B) DATA MODEL AND ONTOLOGY

In this phase, we started with establishing a controlled vocabulary and taxonomy coherent with the categories of Canon Law that historically defined ecclesiastical territorial entities and the changes that could affect a diocese in the course of its existence. This led to the development of a prototype ontology model, *Orbis Dioecesium (OrDi)*, based on CIDOC-CRM with extensions [1]. As the ontology model offers the structuring process for importing data into LOD-bases such as Wikibase or triple-store, this has enabled the transformation of tabular CSV metadata into a structured RDF format.

C) FROM THE DATA MODEL TO WIKIBASE

This part outlines, from a methodological point of view, the challenges of mapping the ontology model described above onto the data model of Wikibase using the data elements within Wikibase: items and properties. The formal modelling process entailed alignment with the Wikibase data model, which was tailored to meet our specific requirements. The first challenge was how to represent the legal-historical changes in the development of the Catholic dioceses, moving away from the traditional descriptive modelling of the data and implementing the historical continuity through an event-based approach. Complexity arises because an event can have multiple causes and effects, precede and follow other events, have start and end dates, and exist within a particular time period. This increases when events are linked to related sequential events, such as the union of two or more dioceses creating a new diocesan organisation at the same time. In the OrDi ontology model, we conceptualised events as a sub-concept of "change" and categorised them according to their nature into sub-classes in the ontology model, OrDi. These were, on the one hand, mapped to the items 'Governance change' (Q16), 'Status change' (Q30), and 'Territorial change' (Q31) in Wikibase, while connecting them to instances of each type.

[5] This value was obtained from the number of existing Catholic dioceses, metropolitan archdioceses, and archdioceses listed on the website GCatholic.org. https://gcatholic.org/dioceses/types.htm/.
[6] These countries denote the states in which the diocesan seats are presently located.

On the other hand, we established a framework for defining properties that capture the temporal sequence of events. Properties such as “follows” (P7) and “followed by” (P8) capture the chronological sequence of events, with guidelines formulated to ensure clarity in the expression of the sequence. Causality within events is systematically addressed through properties such as "caused by" (P13) and "causes" (P14). Semantic representations such as “is erected by” (P18) and “is elevated with higher status” (P40) ensure a standardised representation of the impact of events on entities. Furthermore, the statement that a Catholic diocese was involved in a particular event is represented by the property “participated in” (P29). “In the form of” (P30) is a complementary attribute that can be used to further describe the form of the event.

Second, modelling “Catholic diocese” (Q22) as a sub-concept of “Political or Administrative Entity” (Q18) required a hierarchical approach. Each distinct type of diocese is methodically represented through the creation of specific items, such as “Archdiocese” (Q24), “Metropolis” (Q25), and “Patriarchate” (Q28). In addition, to convey the hierarchical relationships within the church structure, each type of “Ecclesiastical administrative relationship” (Q53) generated items such as “Immediate subiecta” (Q55) and “Suffragan” (Q119) to clearly express whether they are/were directly subordinate to the Holy See or to a single ecclesiastical province.

In addition, when a diocese requires a successor because of a change in status, such as elevation, union, or titular, the properties “replaces” (P11) or “replaced by” (P12) are used to indicate historical continuity.

Third, the categorisation of temporal data involves dealing with events with certain and uncertain dates. In the first case, a property such as “point in time” (P15) is used to specify the exact date, while in the second case additional properties such as “period of occurrence” (P47), “earliest date" (P58) and “latest date” (P59) assign a specific time range. In addition, “EDTF time” (P65) is intended to standardise the representation of ambiguous and uncertain time information.

Fourth, a semantic approach to representing links between events and supporting primary sources was needed, as historical change is documented by “Primary sources” (Q58), such as papal and episcopal decisions, decrees and concordats between secular states and the Holy See. Such primary sources required a semantic approach to represent the links between events and sources, which was facilitated by properties such as “is mentioned” (P42).

Lastly, in the pursuit of interoperability in line with FAIR principles, additional properties were designed to make resources accessible and identifiable through external authority records. Each Uniform Resource Locator (URL) is systematically parsed via a “formatter URL” (P23), contributing to a consistent method for linking and accessing external resources such as *Wikidata* or *VIAF*. Properties such as “equivalent class” (P62) and “equivalent property” (P63) facilitate seamless integration and alignment between the Wikibase data model and external ontological frameworks, improving overall data interoperability.

3. DATA WORKFLOW

Data workflow proceeded as shown in Figure 2: the dataset, organised according to the criterion of the dioceses' Latin names in alphabetical order was saved as a single CSV file and then loaded into OpenRefine (version 3.6.2) for data enrichment and cleaning and was reconciled using the built-in reconciliation service (wikidata). This allowed the dioceses that were later created as instances of Diocese of the Catholic Church (Q22) to be interlinked with other web resources such as Catholic Hierarchy[7], GCatholic.org[8] and BeWeb[9]. The data import pipeline is built in Python using WikibaseIntegrator[10]. The import script iterates through the rows of the CSV file and creates or updates dioceses, events, and bibliographic information depending on whether the item in the CSV already exists in the Wikibase. This process allowed us to automatically import a large number of items into Wikibase and create them. Figure 3 shows the item Parisiensis (Archdiocese of Paris) as it appears on the frontend of the platform.

[7] Catholic-Hierarchy: Its Bishops and Dioceses, Current and Past. https://www.catholic-hierarchy.org/
[8] GCatholic.org. https://gcatholic.org/
[9] BeWeB | Portale Dei Beni Culturali Ecclesiastici. https://beweb.chiesacattolica.it/
[10] WikibaseIntegrator 0.12.4 Documentation. https://wikibaseintegrator.readthedocs.io/en/stable/index.html/

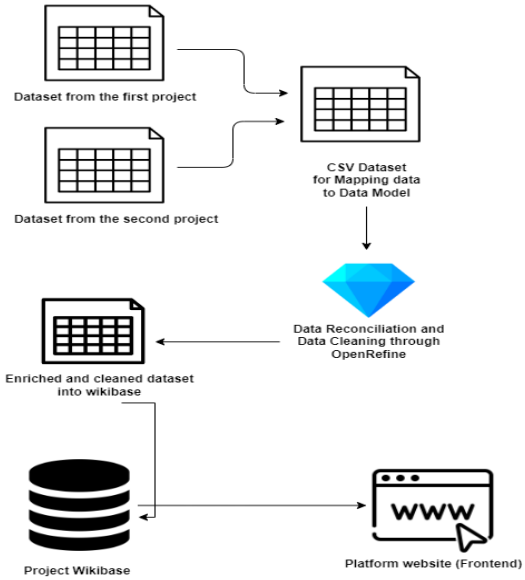


Figure 2. Overall workflow for data collection, cleaning, enrichment, and publication in our platform

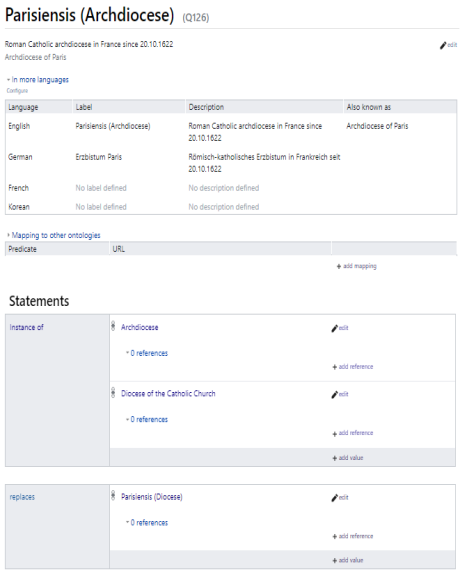


Figure 3. Parisiensis (Archdiocese) created in the project

4. CONCLUSIONS AND OUTLOOK

This paper presents a platform that provides systematically organised information on the legal-historical evolution of dioceses of the Catholic Church from its origins to the present day using the prototype ontology model and knowledge base for storing and publishing authority data according to FAIR Principles. We foresee two main uses for our platform by the scientific community and general public. First, it may be used as an authority record by other research projects interested in providing their data with a methodologically robust framework for representing the nature, the relationships, and the legal changes of ecclesiastical entities throughout history. Second, several quantitative research questions about the legal history of Catholic dioceses could be answered quantitatively by querying the platform's SPARQL endpoint. For example, a user could retrieve information about legal-historical changes that took place at a specific period and the documents that contain this information.

For the future, we plan to implement mapping of the religious geography through visualisations of the changes, which will allow us to provide users with a query to see how the borders of each diocese have changed over time. We will also be gradually updating the dataset by adding those Catholic dioceses missing to the initial database. These prototypes will allow us and others to gain experience with the data we have obtained so far and help guide future development.

REFERENCES

[1] Albani, Benedetta, Alexandra Anokhina, and Yohan Park. 'From the Secret Archive to Open and Fair Access. Ways of Modelling Legal Ecclesiastical Data from the XVI and XVII Centuries', 151. Trier, Luxemburg, 2023.

[2] Battelli, Giuseppe. 'Gli Studi Sui Vescovi e Le Diocesi Del Nord-Italia Tra Cinquecento e Novecento. Panorama Storiografico Dell'ultimo Secolo'. *Rivista di Storia e Letteratura religiosa* 28 (1992): 391–426.

[3] Beretta, Francesco. 'Research Data Interoperability and Foundational Ontologies: An Ecosystem of CIDOC CRMextensions for the Humanities and Social Sciences'. Nancy, France, 2022. https://doi.org/10.5281/zenodo.7014340

[4] Bravo Ugarte, José. *Diócesis y Obispos de La Iglesia Mexicana, 1519-1965. Con un Apéndice de los representantes de La S. Sede en México y viceversa*. México D. F.: Editorial Jus, 1965.

[5] Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Veninata Chiara. 'ArCo: The Italian Cultural Heritage Knowledge Graph'. In *The Semantic Web – ISWC 2019*, edited by Chiara Ghidini, Hartig Olaf, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, 36–52. Cham: Springer International Publishing, 2019.

[6] Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 'Pattern-Based Design Applied to Cultural Heritage Knowledge Graphs'. *Semantic Web* 12, no. 2 (2021): 313–57. https://doi.org/10.3233/SW-200422

[7] De Sandre, Gasparini, Giuseppina Antonio Rigon, Francesco Giovanni Battista Trolese, and Gian Maria Varanini. '*Vescovi e diocesi in Italia Dal XIV Alla Metà Del XVI Secolo. Atti del VII Convegno di Storia della Chiesa in Italia*', Vol. 1990. Roma: Herder, 1987.

[8] Diefenbach, Dennis, Max De Wilde, and Samantha Alipio. 'Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph'. Edited by Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani. *The Semantic Web – ISWC 2021*, 2021, 631–47. https://doi.org/10.1007/978-3-030-88361-4_37

[9] Doerr, Martin. 'The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata'. *AI Magazine* 24, no. 3 (September 2003): 75–92. https://doi.org/10.1609/aimag.v24i3.1720

[10] Echeverria, Lamberto. *Episcopologio Español Contemporáneo (1868-1985). Datos Biográficos y Genealogía Espiritual de Los 585 Obispos Nacidos En España Entre El I. de Enero de 1868 y El 31 de Diciembre de 1985*. Acta Salmanticensia. Derecho. Salamanca: Universidad de Salamanca, 1986.

[11] Hage, Willem Robert van, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 'Design and Use of the Simple Event Model (SEM)'. *Journal of Web Semantics* 9, no. 2 (2011): 128–36. https://doi.org/10.1016/j.websem.2011.03.003

[12] *Historia de las diócesis españolas*. Madrid: Biblioteca de Autores Cristianos, 2002.

[13] Jarry, Eugène, and Jean-Rémy Palanque. *Histoire des diocèses de France*. Paris: Letouzey & Ané, 1967.

[14] Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke Van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank Van Harmelen. 'Semantic Technologies for Historical Research: A Survey'. *Semantic Web* 6, no. 6 (2015): 539–64. https://doi.org/10.3233/SW-140158

[15] Noy, Natasha, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 'Industry-Scale Knowledge Graphs: Lessons and Challenges'. *Commun. ACM* 62, no. 8 (2019): 36–43. https://doi.org/10.1145/3331166

[16] Oviedo Cavada, Carlos, and Marciano Barrios Valdés. *Episcopologio Chileno 1561-1815*. Santiago de Chile: Ediciones Universidad Católica de Chile, 1992.

[17] Plongeron, Bernard, and Jean-Rémy Palanque. *Histoire des diocèses de France. Nouvelle Série*. Paris: Editions Beauchesne, 1974.

[18] Schöch, Christof, Maria Hinzmann, Julia Röttgermann, Katharina Dietz, and Anne Klee. 'Smart Modelling for Literary History'. *International Journal of Humanities and Arts Computing* 16, no. 1 (2022): 78–93. https://doi.org/10.3366/ijhac.2022.0278

[19] Shimizu, Cogan, Pascal Hitzler, Seila Gonzalez-Estrecha, Jeff Goeke-Smith, Dean Rehberger, Catherine Foley, Alicia Sheill, et al. *The Wikibase Approach to the Enslaved.Org Hub Knowledge Graph*. Cham: Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-47243-5_23

Per l'interoperabilità e la sostenibilità delle risorse digitali dantesche: il progetto LiDa

Cesare Concordia[1], Gaia Tomazzoli[2], Nicola Aloia[3], Carlo Meghini[4], Luca Trupiano[5]
[1]CNR Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Italia - cesare.concordia@isti.cnr.it
[2] Sapienza Università di Roma, Italia - gaia.tomazzoli@uniroma1.it
[3]CNR Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Italia - nicola.aloia@isti.cnr.it
[4]CNR Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Italia - carlo.meghini@isti.cnr.it
[5]CNR Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Italia - luca.trupiano@isti.cnr.it

ABSTRACT
In questo contributo presentiamo LiDa (Linking Dante), un progetto che mira a promuovere una maggiore interoperabilità e sostenibilità dei dati sulle opere dantesche raccolti in progetti precedenti tramite l'utilizzo di tecnologie e linguaggi del Web Semantico. Dopo un'introduzione in cui definiamo gli scopi del progetto (§1), facciamo cenno ai limiti con cui si scontra l'annotazione in dei dati linguistici e alle potenzialità di una loro implementazione in RDF (§2); introduciamo poi gli elementi fondamentali della nuova ontologia che abbiamo elaborato a tal scopo (§3), e descriviamo la procedura con cui abbiamo realizzato un nuovo grafo di conoscenza che colleghi il testo della *Commedia* alle risorse linguistiche che lo descrivono (§4); infine, presentiamo le modalità di navigazione e interrogazione di tale grafo (§5) e riflettiamo sui risultati raggiunti e sugli sviluppi futuri (§6).

PAROLE CHIAVE
Ontologie; web semantico; Linguistic Linked Open Data; linguistica italiana; letteratura italiana.

1. INTRODUZIONE
Linking Dante (LiDa) è un progetto che ha l'obiettivo di digitalizzare le opere dantesche e la conoscenza a esse relativa collegando precedenti progetti e implementando nuove funzionalità, per mettere a disposizione dell'utente una piattaforma integrata che renda possibili diverse opzioni di navigazione e interrogazione. La biblioteca digitale di LiDa è basata sulle tecnologie e sui linguaggi del Semantic Web: rispetta i principi FAIR (Findable, Accessible, Interoperable, Reusable) e si ispira alle ontologie di riferimento per la rappresentazione dei dati linguistici e del patrimonio culturale; il testo di Dante diventa così un nodo inserito in una rete dove sono connesse varie risorse che permettono di approfondire le sue caratteristiche linguistiche, semantiche e intertestuali. Per raggiungere questo scopo, LiDa ha tradotto la conoscenza relativa al testo dantesco che si trova attualmente disseminata in diversi strumenti digitali e cartacei in un grafo strutturato secondo una logica calcolabile. In particolare, LiDa integra la conoscenza raccolta dai progetti DanteSearch [9], DanteSources [1], Hypermedia Dante Network [2] e MOnt [5]; in questo contributo ci concentriamo sul dataset di DanteSearch (§2), che abbiamo modellato tramite una nuova ontologia (§3) e collegato al testo della *Commedia* grazie a un algoritmo (§4), rendendolo navigabile e permettendo interrogazioni complesse (§5).

2. DA XML A RDF
Gli studi danteschi sono da diversi decenni all'avanguardia nel campo delle digital humanities (per una rassegna cf. [4]); uno dei progetti più pionieristici in tal senso è stato DanteSearch, una piattaforma digitale che permette la navigazione e l'interrogazione dell'intero corpus delle opere volgari e latine di Dante, che sono state lemmatizzate e annotate per descriverne la morfologia e la sintassi. Il dataset di DanteSearch è costituito da un insieme di file XML annotati secondo la codifica TEI[1]; ciascun file riporta una sola tipologia di annotazione: sintattica oppure morfologica; all'interno dell'annotazione sintattica sono incluse anche alcune informazioni sui dialoghi della *Commedia* (il tipo di discorso – diretto, indiretto o pensato – e il locutore). Una delle funzionalità mancanti nel sistema DanteSearch, discussa dettagliatamente dal coordinatore del progetto [9: 609], è la possibilità di interrogare i testi del corpus definendo contemporaneamente filtri sulla sintassi e sulla morfologia, e questo perché i file con annotazioni di tipo diverso hanno strutture XML/TEI non omogenee tra loro (vd. Fig. 3). Per implementare questa funzionalità in DanteSearch si potrebbero prospettare diverse soluzioni: si potrebbe ad esempio definire una nuova struttura XML/TEI, molto complessa, e mappare in essa le strutture delle diverse annotazioni, oppure creare un indice esterno con un tool specifico, o ancora implementare un modulo software che esegua l'integrazione a runtime, elaborando il risultato di query separate. Ciascuna di queste soluzioni rimane problematica a causa del formato di codifica adottato: XML nasce come formato per la serializzazione dei dati, cioè per la

[1] https://dantesearch.dantenetwork.it/download.jsp

codifica delle informazioni, e per le sue caratteristiche permette di rappresentare facilmente documenti con strutture ad albero, ma richiede soluzioni specifiche per documenti in cui le associazioni tra le parti siano più complesse, come nel caso di un testo annotato. Una delle soluzioni più usate per annotare testi è la codifica XML/TEI, considerata lo «standard internazionale [...] imprescindibile per qualunque impresa di filologia digitale» [9: 586] e adottata, come si diceva, in DanteSearch. Benché le linee guida della TEI definiscano un modulo lightweight per l'annotazione morfologica o sintattica[2] – e non siano dunque incompatibili con l'annotazione di un testo dalla struttura più complessa – i problemi a cui abbiamo accennato ne risultano solo parzialmente risolti, come dimostra il fatto che gli annotatori di DanteSearch hanno elaborato un modello di codifica ad hoc, limitando l'interoperabilità della marcatura con altri corpora.

In anni recenti si sono diffusi progetti che mirano invece a favorire l'interoperabilità tra grandi corpora di testi, e dunque la produzione di Linguistic Linked Open Data (LLOD), grazie ad annotazioni linguistiche fondate su ontologie come Ontolex[3], elaborate in una prospettiva che non guarda primariamente al testo, ma al linguaggio più in generale, e dunque più attenta alle sue varie componenti e alle relazioni tra queste. Un'iniziativa importante in questa direzione è LiLa (Linking Latin), che si propone di connettere varie risorse linguistiche relative al latino per renderle interoperabili, e che comprende corpora annotati, lemma-banks, tree-banks per l'annotazione sintattica e altri strumenti per il NLP. L'elemento centrale dell'architettura di LiLa è però il lemma, il che permette un buon compromesso tra fattibilità e granularità [6: 75]; il nostro approccio, invece, assegna una maggior centralità al testo, inteso come stratificazione di strutture di significato che intrattengono diverse relazioni paradigmatiche e sintagmatiche.

Il progetto LiDa ha tra i suoi obiettivi quello di rendere le opere dantesche annotate interoperabili con altri corpora. Per fare questo abbiamo in primo luogo elaborato un'ontologia che descrive il testo letterario nella sua complessa relazione di strutture di significato relative rispettivamente alla partizione dell'opera, alla morfologia e alla sintassi, e che permette di connettere ogni elemento di tali strutture a modelli e vocabolari standard. Per rappresentare le entità individuate abbiamo scelto di usare RDF, un formalismo per la definizione di modelli di dati basato sui principi del Semantic Web, che permette di creare grafi di conoscenza (knowledge graphs), cioè strutture dati in cui le informazioni sono organizzate secondo regole definite da ontologie formali, e sulle quali è possibile applicare ragionatori automatici per inferire nuova conoscenza. Sui grafi di conoscenza è possibile fare ricerche semantiche e tutte le entità sono identificate da un IRI, che le rende disponibili per essere referenziate, e dunque riutilizzate, come LLOD.

3. CONCETTI E ONTOLOGIE

Il progetto LiDa, nello specifico, si basa su un'ontologia applicativa, espressa in OWL 2 DL e chiamata Ontology of Literary Resources (OLiRes), che a sua volta si fonda sul CIDOC CRM[4] e su Ontolex. OLiRes rappresenta il testo della *Commedia* di Dante secondo la sua struttura testuale, morfologica e sintattica, ed è interoperabile con le ontologie HDN (cf. [2]), MOnt (cf. [5]), ORL (Ontologia delle Risorse Lessicali) e SyntIt (la nostra ontologia per la sintassi dell'italiano), che integrano ulteriore conoscenza rispettivamente sulla dimensione intertestuale codificata dal secolare commento, sulle metafore e su lessico, morfologia e sintassi della *Commedia*. OLiRes rappresenta la conoscenza sul testo delle opere letterarie a tre livelli: 1. il *livello dell'occorrenza*, che rappresenta le caratteristiche del testo nella sua concreta realizzazione, compresa la sua struttura (nel caso della *Commedia*, per esempio, la sua divisione in cantiche e canti); 2. il *livello linguistico*, che codifica gli aspetti linguistici del testo, come la sua morfologia e la sua sintassi; 3. il *livello concettuale*, che rappresenta la componente semantica e concettuale del testo, comprese le relazioni tra le forme che occorrono nel testo e i loro significati o tra il veicolo e il tenore di una metafora.

Questi tre livelli sono legati tra di loro da relazioni di *istanziazione* o di *occorrenza*: ogni entità linguistica è un'istanza della corrispondente entità concettuale e si manifesta in un'occorrenza; così, per fare un esempio, la forma 'cammin' è un'istanza del lemma 'cammino', che a sua volta rimanda al concetto di 'cammino', e occorre nel primo verso del primo canto della *Commedia*, in quarta posizione («Nel mezzo del *cammin* di nostra vita», *Inf.* I, 1). Le entità che fanno parte di uno stesso livello – dell'occorrenza, linguistico o concettuale – sono legate tra di loro da una relazione di *composizione ordinata*: una cantica è composta da una sequenza ordinata di canti, un periodo è composto da una sequenza ordinata di frasi. La relazione di composizione si lega a quella di *localizzazione*: per rappresentare le diverse entità che compongono il testo e per permetterne il reperimento tramite interrogazioni, l'ontologia rappresenta il frammento (o i frammenti) di testo in cui occorre ciascuna entità, specificandone le coordinate di inizio e fine all'interno del testo dell'opera, visto come uno spazio monodimensionale continuo e totalmente ordinato, pur con alcuni distinguo. Una determinata forma flessa, infatti,

[2] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html#AILALW
[3] https://www.w3.org/community/ontolex/wiki/Main_Page
[4] https://www.cidoc-crm.org/

occorre sempre in un unico frammento di lunghezza 1, un determinato periodo occorre sempre in un frammento di lunghezza (almeno) 1, mentre una frase può occorrere in più frammenti se nel periodo si trova intervallata da altre frasi.
All'interno della nostra ontologia, ogni entità, qualunque sia il livello a cui appartiene, è rappresentata dalla classe `olires:Entity`, sottoclasse di `owl:Thing`; poiché nello standard che abbiamo preso a riferimento, il CIDOC CRM, esiste solo la classe `ecrm:E33_Linguistic_Object`[5], che non distingue tra entità linguistiche e frammenti testuali, abbiamo introdotto le due sottoclassi `olires:SyntacticEntity`, avente come istanze le entità linguistiche, e `olires:TextFragment`, avente come istanze le occorrenze (vd. Fig. 2). Per rappresentare le proprietà di istanziazione e occorrenza sopra descritte abbiamo introdotto le proprietà `olires:instanceOf`, che lega ogni istanza del livello linguistico alla corrispondente istanza del livello concettuale, e `olires:occurrenceOf`, che lega ogni occorrenza alla corrispondente istanza del livello linguistico; non è stato possibile ricondurre queste due proprietà ad alcuna proprietà del CRM. Per rappresentare la proprietà di composizione, invece, abbiamo usato la proprietà `ecrm:P148_has_component`, che lega le istanze di `ecrm:E89_Propositional_Object` tra di loro e si applica quindi sia alle entità linguistiche che alle occorrenze, essendo entrambe istanze di `ecrm:E33_Linguistic_Object`, che è sottoclasse di `ecrm:E89_Propositional_Object`.

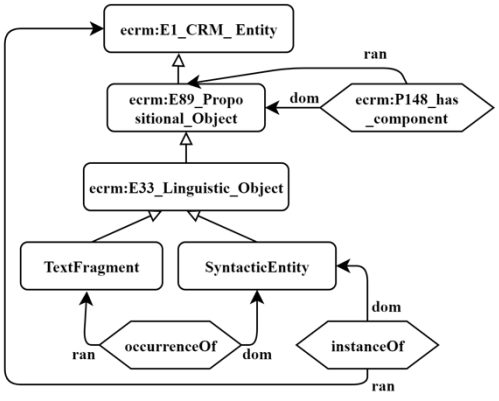


Figura 1. Entità linguistiche e testo

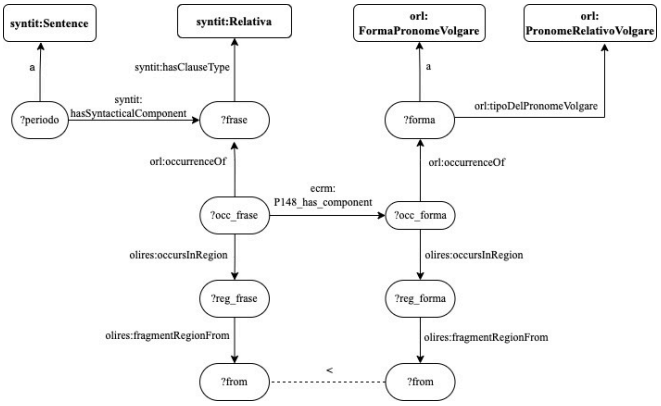


Figura 2. Es. di filtro SPARQL sul grafo della Commedia

Tuttavia, la proprietà `ecrm:P148_has_component` non può essere usata in alcun assioma che vincoli la proprietà a specifiche classi (ad esempio, l'assioma che stabilisce che le componenti di un'entità sintattica sono solo entità sintattiche, e viceversa) perché è dichiarata transitiva, quindi è una proprietà *composite*[6] su cui non possono essere espressi assiomi di cardinalità, pena la violazione della *Restriction on Simple Roles*. Per questo abbiamo introdotto quattro sottoproprietà intransitive di `ecrm:P148_has_component: olires:hasStructuralComponent` per la composizione delle entità strutturali; `olires:hasSyntacticComponent` per la composizione delle entità lessicali e sintattiche; `hasTenor` e `hasVehicle` per la composizione delle metafore.
Come abbiamo ricordato, l'ontologia di LiDa è OLiRes, la quale si appoggia a diverse ontologie di dominio, configurandosi come un sistema composito; le due classi `olires:SyntacticEntity` e `olires:TextFragment` giocano un ruolo fondamentale nell'integrazione delle ontologie che assiomatizzano i diversi fenomeni linguistici che formano tale sistema composito. In particolare, (vd. Fig. 1) l'ontologia Ontolex, usata per la morfologia, si lega a OLiRes tramite gli assiomi che stabiliscono che `ontolex:Form`, la classe delle forme, è sottoclasse di `olires:SyntacticEntity` e che `ontolex:lexicalForm`, la proprietà che lega una forma al suo lemma, è sottoproprietà di `ecrm:P148_has_component`; per contro, la classe `olires:FormOccurrence`, che ha come istanze le occorrenze delle forme, è sottoclasse di `olires:TextFragment`; (2) l'ontologia SyntIt, usata per la sintassi dell'italiano, si lega a OLiRes tramite gli assiomi che stabiliscono che `syntit:Sentence`, la classe dei periodi, è sottoclasse di `olires:SyntacticEntity`, mentre `syntit:SentenceOccurrence`, la classe delle occorrenze dei periodi, è sottoclasse di `olires:TextFragment`; questa classificazione si propaga poi alle relative sottoclassi, quella delle frasi

[5] Secondo la Scope Note, la classe E33_Linguistic_Object del CIDOC CRM «comprises identifiable expressions in natural language or languages. Instances of E33 Linguistic Object can be expressed in many ways: e.g., as written texts, recorded speech or sign language».
[6] Secondo la terminologia introdotta nella sezione 11 (Global Restrictions on Axioms in OWL 2 DL) della specifica di OWL 2 DL (https://www.w3.org/TR/owl2-syntax/).

(`syntit:Clause`) e relative occorrenze (`syntit:ClauseOccurrence`), quella dei sintagmi (`syntit:Syntagm`) e relative occorrenze (`syntit:SyntagmOccurrence`); (3) l'ontologia HDN, usata per le relazioni intertestuali, si lega a OLiRes tramite gli assiomi che stabiliscono che `hdn:Reference`, la classe più generale dei riferimenti intertestuali, è sottoclasse di `olires:SyntacticEntity`, mentre `hdn:ReferenceOccurrence`, la classe più generale delle occorrenze dei riferimenti intertestuali, è sottoclasse di `olires:TextFragment`; (4) l'ontologia MOnt, usata per le metafore, si lega a OLiRes tramite gli assiomi che stabiliscono che `mont:LinguisticMetaphor`, la classe delle metafore linguistiche, è sottoclasse di `olires:SyntacticEntity`, mentre `mont:MetaphorOccurrence`, la classe delle occorrenze di metafore, è sottoclasse di `olires:TextFragment`. In sostanza, `olires:SyntacticEntity` e `olires:TextFragment` agiscono da elementi cardine nell'articolazione delle ontologie che concorrono a formare la base logica di LiDa, lasciando alle ontologie specifiche appena nominate il compito di rappresentare i rispettivi domini. Questo ci permette di osservare che lo sviluppo di un'ontologia applicativa non può prescindere né dal riferimento a ontologie di dominio, né dal riferimento a un'ontologia top – nel caso di LiDa il CRM e Ontolex, che forniscono i concetti di base, con in testa la classe `ecrm:E33_Linguistic_Object`, di cui `olires:SyntacticEntity` e `olires:TextFragment` sono sottoclassi.

4. REALIZZAZIONE DEL GRAFO DI CONOSCENZA DELLA *COMMEDIA*

Una volta definita l'ontologia OLiRes, il grafo di conoscenza relativo è stato realizzato con un software che estrae ed elabora il testo e le annotazioni presenti nei file XML/TEI con le annotazioni morfologiche e sintattiche pubblicate in DanteSearch. Nei file con le annotazioni morfologiche della *Commedia*, uno per ogni cantica, ogni forma è annotata con un elemento `<LM>`. L'elemento ha due attributi, `@lemma` e `@catg`, i cui valori sono rispettivamente la forma canonica corrispondente alla forma annotata e una codifica delle caratteristiche grammaticali/morfologiche della forma annotata.

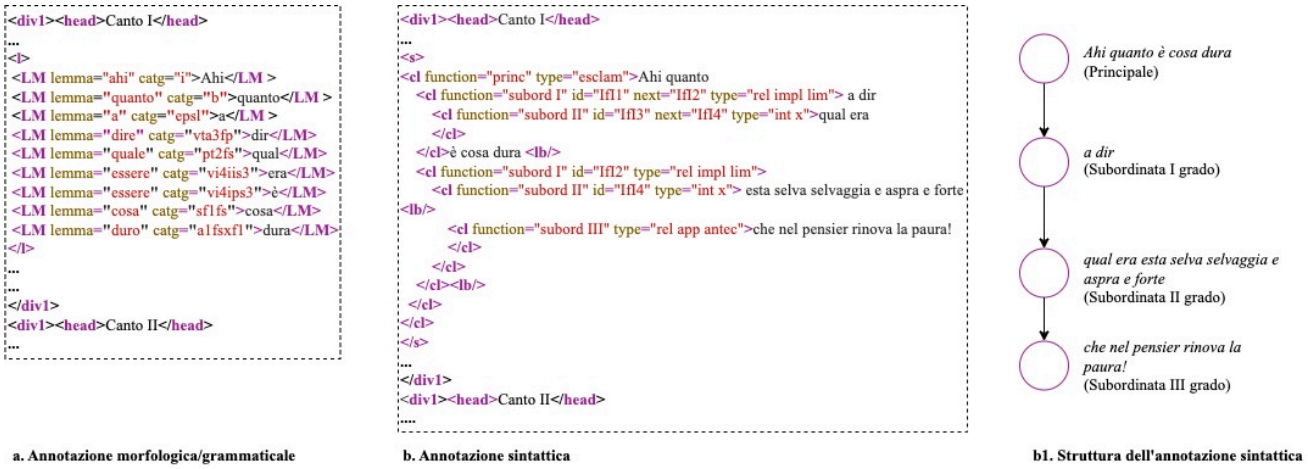


Figura 3. *Esempi di annotazioni della* Commedia *in DanteSearch*

Gli elementi `<div1>` delimitano le forme che appartengono a ciascun canto e gli elementi `<l>` delimitano le forme appartenenti ai versi (vd. Fig. 3a: annotazione morfologica/grammaticale di *Inf.* I, 3). Il software realizzato esegue un parsing dei tre file, processa le annotazioni e genera le entità del livello linguistico e del livello dell'occorrenza per ciascuna delle forme, inserendole nel grafo di conoscenza. L'annotazione sintattica della *Commedia* realizzata da DanteSearch è basata sull'analisi del periodo [3]: ogni canto è suddiviso in periodi, a loro volta suddivisi in frasi; i frammenti di testo che compongono una frase possono essere non consecutivi. Nel file di DanteSearch con le annotazioni sintattiche i periodi sono identificati dall'elemento TEI `<s>` (s-unit) e le frasi dall'elemento `<cl>` (clause); entrambi gli elementi possono avere l'attributo `@id` che identifica l'annotazione e l'attributo `@next` (di tipo IDREF), utilizzato per collegare in un'unica aggregazione elementi che annotano frammenti di testo non consecutivi. Il tipo sintattico[7] e la funzione sintattica[8] di ciascuna frase sono riportate negli attributi `@type` e `@function` dell'elemento `<cl>` che la annota. Le informazioni sui gradi di subordinazione tra le frasi di un periodo sono codificate sia nelle annotazioni sia nell'annidamento degli elementi

[7] Etichette sintetiche che rappresentano categorie desunte dalla Grande Grammatica Italiana di Consultazione [8].
[8] Possibili valori sono: principale, coordinata a una principale, subordinata (I-VII grado), coordinata a una subordinata (I-V grado), parentetica, coordinata a una parentetica, pseudo-coordinata.

<cl> (vd. Fig. 3b: annotazione sintattica di *Inf.* I, 3-6). Di seguito l'algoritmo per la creazione delle entità del grafo di conoscenza sintattico:

- Trasforma il file con le annotazioni sintattiche in una struttura dati che rappresenta un albero orientato: crea un nodo per ogni elemento TEI di annotazione; fra due nodi c'è un arco se i corrispondenti elementi sono in relazione parent-child (un livello di annidamento) nella struttura del file XML/TEI. Gli elementi che annotano una frase composta da frammenti di testo non consecutivi sono raggruppati in un unico nodo (vd. Fig. 3b1: parte dell'albero relativo al periodo di *Inf.* I, 3-6). La struttura dati ottenuta rappresenta un albero orientato, ed è possibile usare degli algoritmi di visita dell'albero per verificare la correttezza del parsing come pure per ottenere informazioni da aggiungere al grafo di conoscenza. Ad esempio, controllando la matrice delle adiacenze delle funzioni sintattiche delle frasi ottenuta visitando l'albero (vd. Fig. 4) sono state rilevate e corrette delle imprecisioni nelle annotazioni.
- Crea le triple delle entità linguistiche sintattiche visitando i sottoalberi che rappresentano i periodi. Parallelamente genera anche le triple del livello dell'occorrenza: per ciascun periodo/frase calcola i frammenti di testo in cui occorre e costruisce le entità del grafo di conoscenza che le rappresentano.
- Utilizzando le entità del grafo generate nella fase precedente individua le occorrenze delle forme la cui posizione è compresa nei frammenti di testo (`olires:TextFragment`) di ciascun periodo e di ciascuna frase; associa tali occorrenze a quelle dei periodi/frasi (tramite `ecrm:P148_has_component`), realizzando il collegamento tra le annotazioni morfologiche e quelle sintattiche.
- Il grafo di conoscenza della *Commedia* è memorizzato in un triple-store e accessibile tramite query SPARQL. Le condizioni di ricerca possono essere create componendo, tramite gli operatori logici, condizioni morfologiche e/o sintattiche. Inoltre, usando le proprietà dell'occorrenza di una entità linguistica, è possibile definire condizioni anche sui frammenti di testo in cui questa si manifesta.

Come esempio di query consideriamo una ricerca sulla *Commedia* descritta in [9], che può essere formulata così: "cerca le frasi relative in cui sia presente un pronome relativo che non occupi la prima posizione nella frase". In DanteSearch questa ricerca viene eseguita solo in parte con il search engine: per essere completata richiede una fase di controllo manuale del risultato ottenuto con l'engine. Nel grafo creato in LiDa, invece, è possibile eseguire la ricerca con una singola query, definendo nel filtro di ricerca le condizioni sulla posizione nel testo delle occorrenze, sia delle forme sia delle frasi (vd. Fig. 2).

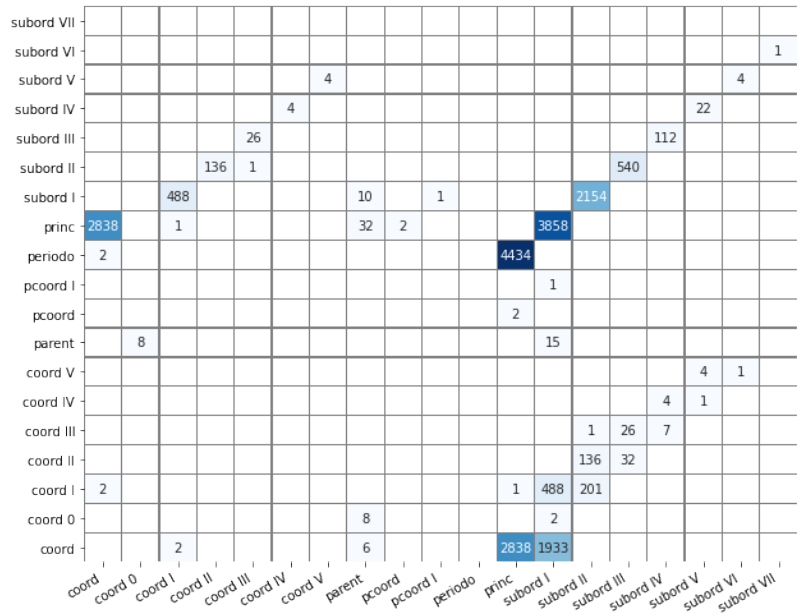


Figura 4. Adiacenze fra funzioni sintattiche

5. VISUALIZZAZIONE E RICERCA

Per permettere, con vari livelli di sofisticazione, l'accesso ai dati del grafo di conoscenza abbiamo realizzato un'interfaccia grafica (GUI) accessibile come Web Application[9]. La GUI fornisce funzionalità di navigazione nel testo della *Commedia* e due modalità di ricerca: semplice e avanzata.

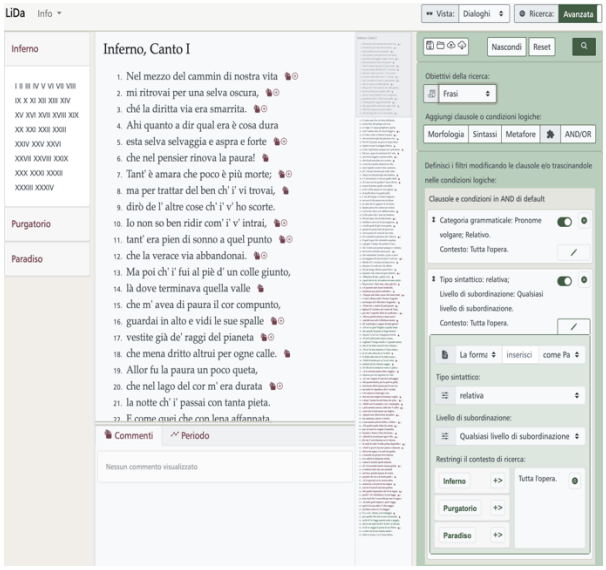


Figura 5. Visualizzazione e ricerca nel testo

La ricerca semplice permette di cercare lemmi, forme o frasi (ma anche prefissi o suffissi) sull'intera opera o su specifiche cantiche e/o canti, esprimendo opzionalmente anche altre condizioni in forma di espressioni regolari. La ricerca avanzata permette di esprimere filtri di ricerca complessi, definendo espressioni booleane i cui operandi possono essere condizioni morfologiche e/o sintattiche. Ricerca e browsing sono visualmente integrate: si può passare da una modalità all'altra senza perdita del contesto, navigando l'opera o esaminando il risultato di una ricerca. Graficamente l'interfaccia è suddivisa in tre aree sovrastate da un menù (vd. Fig. 5): a sinistra si trova un menù di navigazione tra le cantiche e i canti o tra i risultati di una ricerca; l'area di destra presenta gli strumenti per definire le query; l'area centrale contiene la visualizzazione dei dati: vi compaiono il testo del canto selezionato e altre informazioni. Quest'ultima area di visualizzazione è suddivisa orizzontalmente in due zone: nella zona superiore viene visualizzato il canto, con diverse tipologie di visualizzazione, chiamate viste:

- Vista Forme: posizionando il mouse su una forma, vengono visualizzate le informazioni morfologiche e lessicali pertinenti.
- Vista Periodi: tramite simboli e colori vengono delimitati sia i periodi in cui il canto è suddiviso, sia le frasi che compongono ciascun periodo; posizionando il mouse su una frase viene mostrata la sua funzione sintattica.
- Vista Dialoghi: vengono evidenziati i dialoghi (o discorsi) presenti nel canto; per ciascun discorso è possibile visualizzare il parlante e il tipo di discorso (diretto, riportato, pensato).

Nella parte inferiore dell'area di visualizzazione è possibile vedere i commenti relativi alle fonti primarie raccolte dal progetto HDN, facendo un clic sull'icona che compare accanto ai versi per cui esiste almeno un commento, oppure, con un doppio clic su una forma, la rappresentazione grafica, con nodi e archi, della struttura del periodo in cui essa compare.

6. CONCLUSIONI

L'attività di LiDa è ancora in corso: restano da implementare ulteriori funzionalità di ricerca per i commenti relativi alle fonti primarie annotati dal progetto HDN; le metafore di MOnt devono ancora essere integrate nel grafo di conoscenza, e dunque anche la loro interrogazione e visualizzazione tramite la GUI dev'essere implementata. Dato lo scopo di LiDa, che è quello di rendere interoperabili tutti i dati che costituiscono la nostra conoscenza sul testo dantesco, il progetto potrà accogliere in futuro altre risorse: speriamo così di valorizzare il lavoro – passato, presente e futuro – della comunità degli studi danteschi, e di contribuire alla sostenibilità delle risorse digitali attraverso la loro riconversione al paradigma del Web Semantico, con i suoi modelli e linguaggi. L'integrazione tra i dati relativi a tanti progetti differenti non sarebbe stata possibile con una struttura rigida come quella delle annotazioni sintattiche e morfologiche basate sull'XML: sarebbe stato estremamente laborioso, se non impossibile, realizzare una struttura ad albero così stratificata da definire nel testo dantesco

[9] https://lida.dantenetwork.it

decine di migliaia di frammenti di natura e granularità diversa. Il problema non è solo quello tecnico della sovrapposizione di gerarchie (cf. [7]): un approccio che si concentri sul testo e sulla sua espressione linguistica rende evidente la varietà di strutture sintagmatiche e paradigmatiche, e dunque di unità, con cui la lingua costruisce i suoi significati. Le triple RDF costituiscono invece una struttura liquida e seguono un pattern semplice, che definisce una regione di testo e permette la sua annotazione a qualunque livello di granularità; hanno inoltre il grande vantaggio di dotare ogni singola risorsa di un IRI, che la rende disponibile per essere referenziata e dunque riutilizzata come LLOD.

BIBLIOGRAFIA

[1] Bartalesi, Valentina, Carlo Meghini, Daniele Metilli, Mirko Tavoni, and Paola Andriani. 'A Web Application for Exploring Primary Sources: The DanteSources Case Study'. *Digital Scholarship in the Humanities* 33, no. 4 (2018): 705–723.

[2] Bartalesi, Valentina, Nicolò Pratelli, Carlo Meghini, Daniele Metilli, Gaia Tomazzoli, Leyla Livraghi, and Michelangelo Zaccarello. 'A Formal Representation of the Divine Comedy's Primary Sources: The Hypermedia Dante Network Ontology'. *Digital Scholarship in the Humanities* 37, no. 2 (2022): 630–643.

[3] Gigli, Sara. 'La codifica sintattica della Commedia di Dante'. In *Sintassi dell'italiano antico e sintassi di Dante. Atti del seminario di studi (Pisa, 15-16 ottobre 2011)*, edited by Marta D'Amico, 81–96. Ghezzano: Felici, 2015.

[4] Maselli, Matteo. 'Per una rassegna degli strumenti della critica dantesca: dai repertori testuali ai dispositivi digitali'. *Paratesto* 18 (2021): 299–337.

[5] Meghini, Carlo, and Gaia Tomazzoli. '*Per un'ontologia delle metafore nella Commedia di Dante*'. In «Per intelletto umano / e per autoritadi». Il contesto di formazione e diffusione culturale del poema dantesco. Atti del I Convegno HDN (Pisa-Firenze, 29-31 ottobre 2020), edited by Leyla Livraghi and Gaia Tomazzoli, 127–152. Firenze: Franco Cesati, 2022.

[6] Pedonese, Giulia, Flavio M. Cecchini, and Marco Passarotti. 'Linking the Computational Historical Semantics Corpus to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin'. In *Language, Data and Knowledge 2023*, edited by Sara Carvalho, Anas F. Khan, Ana Ostroski Anic, Blerina Spahiu, Jorge Gracia, John P. McCrae, Dagmar Gromann, Barbara Heinisch, and Ana Salgado, 74–85. Lisboa: NOVA FCSH - CLUNL, 2023.

[7] Renear, Allen, Elli Mylonas, and David Durand. Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies, 1993.

[8] Renzi, Lorenzo, Matteo Salvi, and Anna Cardinaletti, eds. *Grande grammatica italiana di consultazione*. Bologna: Il Mulino, 1995.

[9] Tavoni, Mirko. 'DanteSearch: Il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica'. In *Lectura Dantis 2002-2009. Omaggio a Vincenzo Placella per i suoi settanta anni*, edited by Anna Cerbo, II:583–608. Napoli: Università L'Orientale, 2011.

Per un'analisi dei personaggi tra letteratura, filosofia e ontologia applicata

Emilio M. Sanfilippo[1], Gaia Tomazzoli[2], Michele Paolini Paoletti[3],
Jansan Favazzo[4], Roberta Ferrario[5]
[1] CNR ISTC Laboratorio di Ontologia Applicata, Italia - emilio.sanfilippo@cnr.it
[2] Università Sapienza di Roma, Dipartimento di Studi Europei, Italia - gaia.tomazzoli@uniroma1.it
[3]Università degli Studi di Macerata, Dipartimento di Studi Umanistici, Italia - m.paolinipaoletti@unimc.it
[4]Università degli Studi di Macerata, Dipartimento di Studi Umanistici, Italia - jansan.favazzo@hotmail.it
[5]CNR ISTC Laboratorio di Ontologia Applicata, Italia - roberta.ferrario@cnr.it

ABSTRACT

Tanto a livello teorico quanto a livello applicativo, nelle *digital humanities* la rappresentazione della conoscenza si focalizza principalmente sulla modellizzazione dei dati materiali dei prodotti culturali digitalizzati (testi, opere d'arte, ecc.), mettendo in secondo piano la rappresentazione delle *opinioni* degli studiosi su tali prodotti. La parzialità di questo approccio è evidente nel caso della *storia della letteratura* e della *critica letteraria*, discipline in cui le interpretazioni dei testi, con le loro argomentazioni e relazioni reciproche, sono fondamentali anche nell'alimentare dibattiti metodologici. In questo lavoro presentiamo i primi passi di un approccio interdisciplinare pensato per *documentare* e possibilmente *comparare* e *analizzare* le interpretazioni fornite dalla critica letteraria tramite modelli formali sviluppati grazie al coinvolgimento attivo di esperti della disciplina e all'analisi della storia della critica. Alla luce della loro rilevanza e della loro sfaccettata fortuna, abbiamo scelto come casi di studio alcuni personaggi femminili delle opere di Dante e Boccaccio. Dopo aver presentato i principali problemi sollevati dall'analisi dei personaggi letterari e i casi di studio che abbiamo scelto, riflettiamo su come rappresentare le interpretazioni relative ai nostri casi di studio tramite modelli ontologici e sul contributo degli studi filosofici al problema dell'identità dei personaggi di finzione.

PAROLE CHIAVE

History of literature; interpretation; fictional characters; ontology.

1. INTRODUZIONE

Nel contesto delle ontologie formali applicate alla rappresentazione della conoscenza e alla gestione dei dati in ambito umanistico si è soliti concentrarsi su entità di vario tipo, siano esse opere d'arte figurativa, reperti archeologici, documenti storici o altro ancora. Questo approccio ha prodotto grandi quantità di dati, che risultano oggi a disposizione di studiosi, *stakeholders* e studenti per ulteriori ricerche e sviluppi tecnologici. In questo contesto restano tuttavia escluse risorse che potremmo chiamare *osservazionali* o *interpretative*, ossia quelle relative all'espressione del *punto di vista critico*, delle *interpretazioni* elaborate dagli esperti delle varie discipline (storia, critica d'arte, archeologia, ecc.). Se prendiamo il caso che qui ci interessa, ovvero la critica letteraria e la storia della letteratura, le informazioni tipicamente gestite sulle piattaforme digitali si limitano spesso ai metadati relativi alla dimensione materiale dei testi e ai loro contesti di produzione[1]. Una nuova sfida per le *digital humanities* potrebbe dunque essere quella di rappresentare le opinioni dei critici e degli studiosi: cosa sostengono e sulla base di quali fonti, argomentazioni e teorie critiche, ma anche quali relazioni – di derivazione, accordo, disaccordo o altro – sussistono tra le varie opinioni e come tali relazioni sono esplicitate. Un sistema progettato in questa direzione sarà, in linea di principio, capace sia di *documentare* il dibattito critico all'interno di un certo dominio, sia di *analizzare* e *comparare* le varie osservazioni, facendo emergere, ad esempio, la pluralità dei punti di vista critici, le loro possibili relazioni, le ragioni per sostenere o confutare una certa tesi e così via.

Alcune iniziative di ricerca in questa direzione sono state avanzate negli anni 2000 [3] per poi essere accantonate, mentre nei più recenti studi di informatica umanistica stanno riguadagnando un nuovo spazio [1; 2; 7; 22]. Rimangono però ancora da indagare diversi aspetti; ad esempio, come collocare le osservazioni all'interno di un discorso più generale sui modelli usati per la rappresentazione della conoscenza, come gestire osservazioni in reciproco conflitto, come affrontare le loro diversità concettuali, o fino a che grado è possibile rappresentare la struttura di un'argomentazione condotta in linguaggio naturale attraverso metodi formali – per menzionare solo alcuni aspetti.

L'obiettivo di questo contributo è presentare alcuni interrogativi e i primi passi della nostra proposta per un modello ontologico elaborato per documentare dati osservazionali in ambito umanistico attraverso una stretta collaborazione tra

[1] Come esempio, si vedano i dati associati agli esemplari della *Commedia* di Dante su Europeana: https://www.europeana.eu.

studi di critica letteraria, filosofia e ontologia applicata. Per il momento la nostra riflessione si concentra sulla ricerca in ambito letterario e in particolare sulle interpretazioni dei *personaggi*, data la loro importanza all'interno della letteratura in ogni sua epoca e genere, ma auspichiamo che sia generalizzabile e applicabile anche ad altri casi. L'intento è quello di sostenere, mediante l'introduzione di nuove metodologie e strumenti analitici, il lavoro di studiosi e critici nel reperimento, nell'analisi e nel confronto dei dati osservazionali prodotti dalla comunità scientifica, in modo da valorizzarne il lavoro (per l'impostazione di metodo, cf. [10]).

2. I PERSONAGGI LETTERARI E LE LORO INTERPRETAZIONI

I testi sono oggetti complessi, con molteplici livelli di significato irriducibili a una componente oggettiva o formale perché radicati in scelte e posture dell'autore e del lettore [4: 33]. I metodi con cui si può leggere un testo letterario sono, come noto, numerosi e stratificati nel tempo e nello spazio, e le interpretazioni che ne scaturiscono sono tanto diverse quanto diversi sono i lettori. Il nostro modo di leggere i testi è condizionato da un ampio numero di fattori, che vanno dalle nostre competenze linguistiche al canone che ci è stato trasmesso dalle istituzioni, passando per i processi cognitivi e le reazioni emotive sollecitate dalla lettura e per l'interazione tra la condizione socioculturale, le conoscenze situate e i sistemi assiologici del lettore da una parte, e, dall'altra, l'etica e la visione del mondo veicolate dall'opera. Quale che sia il nostro paradigma, la nozione di personaggio, che pure è difficile da definire analiticamente, è al centro della riflessione sulla letteratura almeno fin da Aristotele; dopo una fase di raffreddamento seguita all'esplosione della narratologia nel secolo scorso, negli anni Duemila le teorie sul personaggio hanno cominciato nuovamente a moltiplicarsi, specialmente in ambito anglo-americano, dove il personaggio è stato indagato soprattutto in chiave formale, sociale, cognitiva, analitica, psicoanalitica e semiotica [5]; molti di questi studi hanno messo in rilievo la stretta relazione che esiste tra la caratterizzazione del personaggio e l'interpretazione che il lettore dà di tale caratterizzazione. Per maneggiare meglio la grande quantità e complessità di teorie sul personaggio nate nell'ambito della teoria della letteratura (su cui cf. ad esempio [13]), il nostro lavoro intende concentrarsi su alcuni casi di studio specifici, che ci permettano di combinare l'approccio *top-down*, basato sullo studio di tali teorie, con un approccio *bottom-up*, fondato sulla modellizzazione di alcuni degli aspetti evidenziati da un corpus di saggi critici dedicati ai nostri casi di studio.

Abbiamo scelto di concentrarci su alcuni personaggi femminili di due autori fondamentali per le origini della letteratura italiana: due personaggi danteschi (Beatrice e Francesca) e due personaggi boccacceschi (Fiammetta e Griselda); si tratta di personaggi complessi e affascinanti, oggetto di una ricchissima ricezione interpretativa e creativa e per questo punto d'accesso privilegiato per osservare la stratificazione degli approcci critici e compararli sulla *longue durée*. La scelta è stata dettata da diverse ragioni: Beatrice e Fiammetta occupano uno statuto particolare nei macrotesti dei due autori perché compaiono in diverse opere con funzioni cangianti, e perché la critica si è a lungo interrogata sulla loro reale esistenza e sul loro rapporto biografico rispettivamente con Dante e con Boccaccio. Se Beatrice è stata letta come «figura mitica di una specie di eterno femminino» [17: 14] e come personaggio essenzialmente allegorico, la musa di Boccaccio, Fiammetta, è un personaggio dalle varie funzioni, continuamente in bilico tra finzionalità e verità storica [21]. Francesca, dal canto suo, è stata vista dai contemporanei di Dante come un'adultera da cronaca nera e come una peccatrice da condannare, mentre i lettori di Otto e Novecento l'hanno letta piuttosto come un archetipo della passione amorosa femminile, dibattendo sulle ragioni per cui condannarla o salvarla [18]; Griselda, viceversa, è stata descritta come un più freddo prototipo di fanciulla perseguitata, modello di pazienza e virtù [16]. Anche in virtù di questa componente esemplare, questi due personaggi sono stati oggetto di numerosissime riscritture, oltre che di molte e divergenti interpretazioni.

Più in generale, i personaggi femminili sono una specola interessante per seguire le evoluzioni storiche della critica letteraria, che negli ultimi decenni è stata profondamente rivoluzionata dal femminismo: i personaggi femminili sono condizionati, a ogni grado della loro esistenza e ricezione, dal genere maschile di chi li ha creati e di chi li interpreta, e per questo aprono una serie di questioni su temi come la valutazione morale dei personaggi, la loro esemplarità, i processi di immedesimazione che si attivano in diverse categorie di lettrici e lettori. Proprio in quanto dotati di un «alto grado di letterarietà», che si manifesta nel loro essere archetipi di comportamento e nel loro richiamarsi ad altri personaggi altrettanto archetipici [6: 9], la critica letteraria si sofferma spesso, nelle sue interpretazioni, sulle fonti dei personaggi femminili, oltre che sui loro attributi, sulle azioni che compiono e dunque sulla loro funzione nell'intreccio, sulle loro relazioni con altri personaggi, sui contesti narrativi, emotivi e morali in cui si trovano ad agire.

Ci sono stati alcuni tentativi di creare un'ontologia dei personaggi per applicazioni digitali: Ciotti [4], per esempio, caratterizza i personaggi in relazione ai loro attributi e alle loro funzioni sulla scia di Greimas, e dunque all'interno di una prospettiva narratologica e semiotica che li considera oggetti testuali esistenti in un mondo possibile. Zöllner-Weber [23] combina diverse teorie del personaggio in un'ontologia che permette di distinguere, per esempio, tratti fissi e tratti mutevoli, o tra caratteristiche che al personaggio vengono attribuite dal narratore, da un altro personaggio o da sé stesso; pur introducendo una classe relativa alle caratteristiche indirette, cioè dipendenti dall'interpretazione del lettore, nemmeno

questa ontologia problematizza però a fondo il rapporto tra il personaggio e il suo interprete. Hastings e Schulz [12] compiono un passo ulteriore notando che tutte le entità che fanno parte del dominio di rappresentazione – tra cui i personaggi – sono prodotti culturali, e dunque intrattengono relazioni con il loro creatore, con il testo in cui sono contenuti e con il loro contesto di produzione, ma sono anche rappresentazioni appartenenti a un dominio specifico, quello creato dal testo, nel quale hanno determinate proprietà. Il nostro approccio si concentra invece proprio su quello che altri considerano un semplice filtro, vale a dire sull'interpretazione di varie tipologie di lettori, con un ruolo speciale assegnato a studiose e studiosi di critica letteraria.

3. LE OSSERVAZIONI NEI MODELLI ONTOLOGICI

Come abbiamo visto nella sezione precedente, critici e studiosi esprimono diversi tipi di osservazioni attraverso diversi paradigmi interpretativi e concentrandosi su diversi aspetti. Se ci si pone in una prospettiva diacronica, inoltre, un modello che rappresenti una determinata interpretazione dovrebbe essere generalizzabile anche ad altri casi. Più in generale, se intendiamo sviluppare un sistema informativo per documentare, analizzare e confrontare osservazioni espresse da più studiosi, avremo bisogno di un linguaggio comune per rappresentare le osservazioni in modo uniforme.

Dal punto di vista del modello, la prospettiva proposta si articola sulla base di precedenti studi formali sulla rappresentazione della conoscenza, e in particolare sull'espressione di dati osservazionali in ambito scientifico e ingegneristico [15]. Nonostante gli ambiti di applicazione siano diversi, ci proponiamo di estendere l'approccio di Masolo e colleghi [15], che presenta un carattere di generalità tale da poter essere utilizzato in diversi contesti senza alterarne il contributo sostanziale. In sintesi, l'idea è sviluppare uno o più *linguaggi osservazionali* i cui termini rappresentano *tipi di osservazioni* organizzati tassonomicamente a più livelli e riguardanti uno o più entità. In particolare, un linguaggio osservazionale consiste sia di un *vocabolario controllato*, sviluppato in collaborazione con gli esperti di dominio – studiosi di letteratura e critici letterari, nel nostro caso –, sia di *strumenti logici* per caratterizzare il significato dei termini del vocabolario, *confrontare* le interpretazioni di diversi studiosi, *documentare* (come vedremo) informazioni legate alla *provenienza* delle interpretazioni o agli argomenti a *supporto* di una certa osservazione, per menzionare solo alcuni aspetti. In tal senso un linguaggio osservazionale è più ricco di una terminologia usata per annotare un testo (cf., ad es., [11]), dato che include aspetti formali comunemente non inclusi nelle terminologie per le annotazioni.

Alla luce della loro funzionalità applicativa, distinguiamo tra le *osservazioni di base*, che riguardano solo entità di dominio diverse dalle osservazioni, e *osservazioni complesse*, che riguardano invece le osservazioni stesse. Un esempio tipico delle prime è l'*attribuzione di proprietà* ai personaggi, mentre alcuni casi di osservazioni complesse sono le osservazioni di *asserzione* (o *rifiuto*) e *supporto*, che legano un'osservazione ad altre osservazioni.[2] La rappresentazione delle osservazioni segue sempre lo stesso schema: si tratta di modellare individui (delle specifiche sottoclassi) della classe *Osservazione* insieme alle entità che partecipano nell'osservazione, ossia le entità prese in esame da un interprete.

Prendiamo il caso di Francesca; assumiamo di voler documentare l'osservazione critica secondo la quale Francesca è una donna *adultera*. Dato che l'osservazione riguarda direttamente il personaggio, si tratta di un'osservazione di base che attribuisce a Francesca la proprietà di *essere adultera*. Inoltre, tale osservazione potrebbe essere intesa come un caso specifico di caratterizzazione del personaggio all'interno di uno spazio di valori morali che include altri tipi di osservazioni. In una logica del primo ordine, il tipo d'osservazione sull'"adulterità" può essere resa con la seguente formula[3]:

$$OsservazioneAdultera(x) \rightarrow Osservazione(x) \wedge \exists y\big(ARG_1(x,y)\big) \wedge \bigwedge_{i=2}^{\alpha} \neg \exists y\big(ARG_i(x,y)\big)$$

dove *x* è un'osservazione di tipo *essere adultera* che riguarda l'entità *y* (considerando la formula, il predicato *ARGn* sta per *argomento*; l'ultima condizione del conseguente asserisce quindi che *OsservazioneAdultera* ha una sola entità come argomento). Nel caso di Francesca possiamo scrivere: $OssservazioneAdultera(o) \wedge ARG_1(o, Francesca)$, dove *o* è una costante individuale per una specifica osservazione e *Francesca* è una costante per il personaggio di Francesca nella *Commedia* di Dante. La formula quindi, che può essere semplificata per brevità sintattica in *o=adultera(Francesca)*, asserisce che l'osservazione *o* riguarda *Francesca* e le attribuisce la proprietà di essere *adultera*. Una simile caratterizzazione formale può essere messa in atto per altri tipi di osservazioni di base, incluse quelle che riguardano non una ma più entità di dominio, come le osservazioni di *somiglianza* tra due o più personaggi [20].

[2] Ci limitiamo a riportare solo alcuni aspetti di una prima bozza di proposta formale; per maggiori dettagli cf. [20].

[3] È utile ricordare che i linguaggi maggiormente utilizzati per lo sviluppo di ontologie computazionali, ossia i linguaggi del Semantic Web, sono basati su frammenti decidibili della logica del primo ordine.

Per le osservazioni complesse, che coinvolgono quindi altre osservazioni, quelle di *asserzione* (*rifiuto*) e *supporto* sono particolarmente rilevanti perché permettono di documentare le fonti di un'osservazione e gli argomenti a suo supporto.
Per quanto riguarda l'*asserzione*, si assume che le osservazioni siano sempre espresse in testi accessibili intersoggettivamente. Pertanto, diversamente da modelli come il CRM Argumentation Model [9], le osservazioni per come le intendiamo noi sono informazioni accessibili indipendentemente dagli stati mentali di chi le esprime. Dati osservazionali su questa scia sono esprimibili prendendo in prestito meccanismi formali dalle logiche per l'argomentazione. Ad esempio, utilizzando una rappresentazione sintattica semplificata, potremmo scrivere che *o=ass(txt,adultera(Francesca))*, introducendo l'osservazione *o* secondo cui nel testo *txt* si *asserisce* (*ass*) che Francesca è un'adultera. In tal senso le osservazioni di *asserzione* sono utili per documentare *cosa* viene asserito (o rifiutato) in una certa fonte. È da notare che l'*asserzione* è un'osservazione nel senso che può essere essa stessa asserita o rifiutata: lettori diversi, per esempio, potrebbero essere in disaccordo sull'asserzione per cui Francesca è una figura adultera secondo un certo testo. È pertanto possibile costruire catene di osservazioni per documentare il dibattito; ad esempio *o=ass(txt',ass(txt,adultera(Francesca)))* – ossia il testo *txt'* asserisce che *txt* asserisce che Francesca è adultera.
Le osservazioni di *supporto* rappresentano un tipo di osservazioni utili per documentare altre osservazioni che mettono in luce, potremmo dire, le *premesse* a sostegno di un'osservazione *conclusiva*. In un certo senso, all'interno del ragionamento di uno studioso, un'osservazione di supporto aumenta la plausibilità di un'altra osservazione. Ad esempio, si potrebbe dire che l'osservazione secondo cui Francesca consuma il tradimento con Paolo ai danni del marito Gianciotto fa da supporto all'attribuzione della proprietà *essere adultera* a Francesca. Sempre utilizzando una rappresentazione formale semplificata, potremmo scrivere qualcosa come:

o=sup(tradimento(Francesca,Gianciotto,Paolo),adultera(Francesca))

dove (i) *adultera* rappresenta l'osservazione di base discussa sopra; (ii) *tradimento* è un'osservazione di base di tipo *OsservazioneTradimentoConiugale* con tre entità come argomenti; (iii) *sup* rappresenta invece un'osservazione complessa del tipo *supporto*. Quest'ultima asserisce che l'osservazione sul tradimento fa da supporto (premessa) all'osservazione (conclusione) di "adulterità" del personaggio di Francesca. È poi possibile combinare osservazioni di *asserzione* e *supporto*, come nella seguente formula, secondo cui l'osservazione di supporto è asserita nel testo *txt* di un certo studioso:

o=ass(txt,sup(tradimento(Francesca,Gianciotto,Paolo),adultera(Francesca)))

Dal punto di vista logico, le osservazioni di supporto seguono anch'esse un trattamento formale affine alle logiche per l'argomentazione e *sup* esprime una condizione più debole rispetto all'implicazione materiale della logica classica (cf. [20]). In particolare, in *sup(o,o')*, l'osservazione *o* non esprime una condizione sufficiente per *o'* ma rappresenta – come dicevamo – un'indicazione a favore della plausibilità di *o'* (su questo punto cf. anche [3]).
Da questi primi brevi esempi possiamo trarre alcune considerazioni.
Primo, la teoria logica quantifica sulle osservazioni, che sono entità nel dominio di quantificazione al pari delle altre; è pertanto possibile rappresentare le loro proprietà – incluse, come abbiamo visto, le informazioni su chi le asserisce.
Secondo, ogni osservazione esprime il punto di vista di qualcuno che, sulla base di certe procedure e argomentazioni, decide quali proprietà attribuire alle entità analizzate. Ad esempio, che Francesca sia un'adultera non rappresenta, potremmo dire, un *fatto*; piuttosto, l'osservazione esprime l'*opinione* o intuizione di uno studioso. Nonostante questo, il criterio d'identità delle osservazioni prescinde da chi le asserisce, lasciando aperta la possibilità che più studiosi esprimano le *stesse* osservazioni indipendentemente gli uni dagli altri.
Terzo, un linguaggio osservazionale intende documentare le osservazioni su un testo, ma resta comunque il prodotto degli obiettivi intellettuali di chi produce i dati osservazionali. Potremmo voler documentare, ad esempio, che secondo un lettore Francesca è il personaggio di un'adultera, ma quella di documentare proprio quest'osservazione, magari a discapito di altre, resta una nostra scelta. In questa prospettiva sono necessari almeno due chiarimenti. Innanzi tutto, è necessario che i linguaggi osservazionali siano sviluppati in accordo tra più interpreti, in modo da estrarre informazioni dai testi in modo uniforme e controllato. Inoltre, in un'ottica di documentazione e analisi delle interpretazioni, chi produce i dati osservazionali deve sforzarsi il più possibile di documentare l'interpretazione del testo e non la propria. Questo pone non pochi problemi, in particolare se assumiamo che la documentazione di un'interpretazione può essa stessa essere concepita come un'interpretazione. La nostra strategia si propone di rendere esplicito il rimando a chi produce i dati osservazionali, in modo tale da poter ricostruire la "catena interpretativa". Sul piano metodologico, inoltre, il processo di produzione dei dati osservazionali deve realizzarsi in una forma *collaborativa* per permettere e facilitare lo scambio di opinioni e l'interazione tra gli interpreti (cf., dal punto di vista metodologico, [11]).

Quarto, come già detto, il carattere plurale degli studi letterari presuppone la possibilità, empiricamente verificabile, che per uno stesso testo si diano interpretazioni non compatibili. Il sistema formale deve essere pertanto capace di gestire il *conflitto* tra le osservazioni evitando la generazione di contraddizioni logiche. Nella nostra proposta è possibile documentare l'asserzione o il rifiuto delle osservazioni, come (parzialmente) mostrato sopra, senza per questo contraddirsi. Infine, le osservazioni che qui ci interessano riguardano *personaggi letterari*. Dal punto di vista della rappresentazione si pone dunque a monte il problema di capire *cos'è* un personaggio e come distinguere diversi personaggi, in particolare quando questi sono oggetto di riscrittura in nuovi testi, spesso di altri autori. Per esempio, quando gli studiosi interpretano un testo in cui appare un personaggio chiamato Francesca si confrontano su un *unico* personaggio la cui fisionomia è già fissata o le loro interpretazioni contribuiscono a definirne l'*identità*? Con quali strumenti metodologici possiamo affrontare questo e altri interrogativi? Com'è prassi nello sviluppo delle ontologie, ci sembra importante avvalerci degli studi sul personaggio sviluppati in ambito filosofico: in tal senso la nostra proposta mira a sviluppare un modello utilizzabile in diversi contesti applicativi ma che sia anche fondato dal punto di vista teorico.

Prima di passare alla sezione successiva è doveroso ribadire che la rappresentazione formale delle osservazioni sta ricevendo crescente attenzione nelle *digital humanities*. Ad es., in un recente contributo, Sartini e colleghi [22] propongono un framework di modellazione per documentare dati osservazionali per l'iconografia. Per menzionare una differenza importante rispetto alla nostra proposta, le loro osservazioni sembrano a prima vista dipendere dagli agenti che le esprimono, mentre nel nostro caso hanno una dimensione più astratta per facilitare l'attribuzione di una *stessa* osservazione a diversi studiosi. Nella proposta di Daquino e colleghi [8], invece, viene utilizzata una sintassi basata sul linguaggio RDF del Semantic Web per l'espressione di quelle che gli autori chiamano *congetture*; in questo caso non si discute delle loro condizioni d'identità, ossia se e in che misura dipendono da chi esprime le congetture. A parte queste e altre differenze fondazionali, la nostra proposta include vari strumenti formali non presenti in altri contributi ed ereditati dalle logiche per l'argomentazione per poter *asserire*, *supportare* ma anche *rifiutare* o *andare contro* una certa osservazione, oltre a un primo insieme di meccanismi logici per il confronto delle osservazioni [20].

4. IDENTITÀ DEI PERSONAGGI FITTIZI E INTERPRETAZIONE

Quanto abbiamo detto nella sezione precedente si fonda sull'implicazione che *Francesca* sia una costante, ossia che corrisponda stabilmente a un personaggio della *Commedia* di Dante; per rappresentare quest'entità in modo esplicito e in quanto personaggio dobbiamo affinare la nostra analisi per rimuovere ogni possibile ambiguità e caratterizzare al meglio i termini utilizzati nel modello formale.

Dal punto di vista filosofico, i personaggi letterari sono tipicamente trattati come *oggetti fittizi letterari* [14]. Per condurre una riflessione ontologica sui loro rapporti con gli atti di interpretazione, con le loro fonti, con i loro autori e con il loro contesto storico, è anzitutto necessario riflettere sui *criteri di esistenza* di tali entità (ammesso che esistano) e, soprattutto, sui loro *criteri di identità*. Viceversa, risulta impossibile chiarire cosa si intende quando si parla di *un certo* personaggio fittizio (e non di un altro), dello *stesso* personaggio o di personaggi *distinti*.

I criteri di identità dei personaggi fittizi (o *ficta*) presentano tipicamente la seguente forma: *necessariamente, x e y sono lo stesso fictum se e solo se P* – laddove P può essere sostituito da una caratteristica o da una serie di caratteristiche di *x* e *y* (ad es. avere le stesse e soltanto le stesse proprietà attribuite negli stessi testi); talvolta, invece, ci si limita a condizioni sufficienti o necessarie per l'identità dei *ficta*. Ad ogni modo, i processi di interpretazione possono essere coinvolti nell'identità dei *ficta* in molteplici modi. Occorre però effettuare alcune distinzioni preliminari.

Assumiamo che Francesca sia un personaggio puramente fittizio – o comunque un personaggio fittizio distinto dalla Francesca "reale" cui potrebbe essersi ispirato Dante. Assumiamo di restringere l'indagine agli atti di interpretazione legittimi o corretti, cioè a quelli fondati su e compatibili con il maggior numero di evidenze possibili sui *ficta* (ad esempio, le storie di cui fanno parte, gli autori di tali storie e ogni altro fattore rilevante). Un atto di interpretazione legittimo può essere inteso come un fatto *relazionale*: il *fictum x* è legittimamente interpretato da un interprete come dotato di una certa caratteristica (es. Francesca è legittimamente interpretata da Mario come un'adultera). L'atto di interpretazione deve essere legittimo (o sufficientemente legittimo) allo scopo di evitare atti di interpretazione incompatibili con le migliori evidenze testuali. Ciò non esclude che possano esserci molteplici interpretazioni egualmente legittime, e anche contrastanti tra di loro, ma comunque plausibili. Ad ogni modo, l'ultimo *relatum* di questo fatto relazionale è un ulteriore fatto (es. Francesca è un'adultera o a Francesca è attribuita la proprietà di essere un'adultera nel testo o nei testi rilevanti). Gli atti di interpretazione legittimi possono essere determinati o dipendenti da vari fattori: le informazioni testuali, le intenzioni degli autori, il contesto di produzione, la comprensione da parte di lettori competenti e/o di critici, ecc. Chiamiamo "*fattori di interpretazione*" tutti questi fattori.

Tornando ai criteri di identità dei *ficta* occorre dunque porsi le seguenti domande:

Gli atti di interpretazione possono figurare in P?

Gli atti di interpretazione possono determinare/causare/vincolare ciò che figura in P ma senza figurare in P?

I fattori di interpretazione possono figurare in P?

I fattori di interpretazione possono determinare/causare/vincolare ciò che figura in P ma senza figurare in P?

Senza la pretesa di rispondere compiutamente a questi interrogativi, occorre rilevare quanto segue. Rispetto a (1), gli atti di interpretazione legittimi sembrano anzitutto ridondanti rispetto agli ulteriori fatti che figurano in essi come *relata* e che potrebbero a loro volta figurare legittimamente e direttamente in P. Ad esempio, il fatto che Francesca sia un'adultera potrebbe figurare direttamente in P, e questo renderebbe l'atto di interpretazione legittimo ridondante in quanto costituente di P. Inoltre, quali atti di interpretazione dovrebbero figurare in P? Di quali interpreti? E cosa accadrebbe se, di fatto, non vi fosse alcun interprete a compiere alcun atto di interpretazione legittimo rispetto a un *fictum*? Il *fictum* dovrebbe cessare di avere condizioni di identità, il che è implausibile. Se, in risposta, si asserisce che gli atti di interpretazione coinvolgono *genericamente* un qualche interprete – ma nessuno in particolare – allora gli atti di interpretazione legittimi sembrano godere di un'ontologia peculiare: sono fatti relazionali, ma uno dei loro *relata* non è un'entità specifica, ma *una qualche entità* di un certo tipo (diversamente da quanto accade tipicamente per i fatti relazionali).

Gli stessi problemi possono presentarsi con (3), anche se occorre esaminare le diverse teorie dell'interpretazione, che potrebbero offrire risposte più o meno convincenti a tali problemi (per es. le intenzioni da parte dell'autore originario di identificare un *fictum* potrebbero non essere ridondanti). D'altro canto, rispetto a (2) e (4), sembra che gli atti di interpretazione legittimi e i fattori di interpretazione possano determinare/causare/vincolare ciò che figura in P, senza perciò figurare in P. Ovviamente anche questa opzione è controversa. Ad esempio, è incompatibile con l'idea che i *ficta* siano dotati di condizioni di identità indipendentemente dagli atti di interpretazione legittima e dai fattori di interpretazione.

Considerazioni di questo tipo sono a nostro avviso non solo fondamentali per poter sviluppare un modello ben fondato per la rappresentazione di dati osservazionali in ambito letterario, ma anche utili per approfondire in direzione analitica il lavoro critico stesso. Per fare un esempio, interrogarsi, come si fa abitualmente nella storia della letteratura, sulla *transfinzionalità* dei personaggi letterari, ossia sulle loro possibili relazioni a cavallo tra diverse opere, autori ed epoche, richiede strumenti analitici quali i criteri ontologici d'identità, per poter capire quando ci troviamo di fronte a un unico o a diversi personaggi (possibilmente in relazione di *derivazione* tra loro). Chiaramente criteri di questa natura non possono essere concepiti in modo aprioristico, ma vanno elaborati in dialogo con gli studiosi della disciplina in modo da essere fondati sulle loro teorie e pratiche interpretative. Nelle successive fasi di sviluppo del nostro lavoro sarà quindi fondamentale portare avanti questo tipo di riflessione filosofica in collaborazione con gli esperti di critica letteraria e integrarla con il modello formale sulle osservazioni discusso nella Sezione 3.

5. DISCUSSIONE

Nelle sezioni precedenti abbiamo discusso alcuni aspetti relativi alle interpretazioni dei personaggi per poi vedere, in forma preliminare, come queste possono essere documentate attraverso un sistema formale. La proposta va sviluppata in più direzioni, inclusa una caratterizzazione del personaggio che prenda in considerazione problematiche inerenti la sua *identità* in relazione a più fattori. Come abbiamo sottolineato, il nostro obiettivo è riflettere da un punto di vista interdisciplinare sulle pratiche interpretative in ambito letterario per lo sviluppo di un'ontologia e di una metodologia che possano fare da supporto per la documentazione, il confronto e l'analisi delle interpretazioni attraverso strumenti analitici e formali.

Per quanto riguarda la nostra proposta, è lecito interrogarsi sulle sue potenzialità e sui suoi limiti, nonché sulla sua posizione all'interno di una più ampia riflessione sugli studi umanistici. Per fare solo un esempio, in un saggio di critica letteraria non è importante solo *il contenuto* dell'argomentazione, ma anche il suo stile, espresso in un linguaggio naturale ricco di sfumature; i sistemi formali, invece, spingono inevitabilmente verso l'*astrazione* e la *semplificazione*: ci saranno pertanto osservazioni critiche che non potranno essere rappresentate in un modo che ne riproduca la ricchezza e complessità. A tal proposito, vale la pena enfatizzare che la nostra proposta non intende in alcun modo ridurre il lavoro degli studiosi a uno scheletro formale, né sostituire il lavoro critico con un approccio meramente "meccanico"; intendiamo piuttosto riflettere sulle interazioni tra le varie metodologie e strategie argomentative per mettere a punto dei nuovi strumenti per l'analisi della storia e della teoria della letteratura. La nostra proposta è inoltre diversa dagli studi di *computational literary criticism* [19] poiché nel nostro caso non intendiamo produrre automaticamente interpretazioni basandoci, ad esempio su pattern statistici; si tratta piuttosto di documentare interpretazioni esistenti in modo controllato e poi di implementare tecniche (semi-automatizzate) per la loro analisi. Inoltre, applicando un metodo di rappresentazione formale della conoscenza saremo capaci di documentare la logica dell'argomentazione di un lettore: questo può portarci a capire come si sviluppa un'argomentazione e quali sono le premesse a supporto delle sue tesi, ed eventualmente a individuarne possibili fallacie argomentative. Ovviamente questo richiede in prima battuta l'identificazione di logiche e criteri da utilizzare per la formalizzazione di un'argomentazione letteraria. Tutto ciò può a nostro avviso essere realizzato solo attraverso una stretta

collaborazione con gli studiosi di letteratura che producono e fruiscono tali interpretazioni e attraverso un rinnovato interesse per il dialogo interdisciplinare.

6. RINGRAZIAMENTI

Questo contributo è supportato dal progetto PRIN 2022 PNRR *Make it explicit: Documenting interpretations of literary fictions with conceptual formal models* (MITE) finanziato dall'Unione Europea – Next Generation EU.

BIBLIOGRAFIA

[1] Barabucci, Gioele, Francesca Tomasi, and Fabio Vitali. 'Supporting Complexity and Conjectures in Cultural Heritage Descriptions'. In *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, edited by Andreas Webes et al., 1–12. CEUR 2810, 2020.

[2] Bartalesi, Valentina, Nicolò Pratelli, Carlo Meghini, Daniele Metilli, Gaia Tomazzoli, Leyla Livraghi, and Michelangelo Zaccarello. 'A Formal Representation of the Divine Comedy's Primary Sources: The Hypermedia Dante Network Ontology'. *Digital Scholarship in the Humanities* 37, no. 2 (2022): 630–643.

[3] Benn, Neil, Simon Buckingham Shum, John Domingue, and Clara Mancini. 'Ontological Foundations for Scholarly Debate Mapping Technology'. *Frontiers in AI and Applications* 172 (2008): 61–73.

[4] Ciotti, Fabio. 'Toward a Formal Ontology for Narrative'. *MatLit* 4, no. 1 (2016): 29–44.

[5] Comparini, Alberto. 'Problemi di "personae". Sulla recente teoria del personaggio nel mondo anglofono (2003-2016)'. *Comparatismi* 1 (2016): 267–274.

[6] Crivelli, Tatiana. *Selvagge e angeliche. Personaggi femminili della tradizione letteraria italiana*. Catania: Insula, 2007.

[7] Daquino, Marilena, Valentina Pasqual, and Francesca Tomasi. 'Knowledge Representation of Digital Hermeneutics of Archival and Literary Sources'. *JLIS: Italian Journal of Library, Archives and Information Science = Rivista Italiana Di Biblioteconomia, Archivistica e Scienza dell'informazione*: *11, 3, 2020*, no. 3 (2020): 59–76. https://doi.org/10.4403/jlis.it-12642

[8] Daquino, Marilena, Valentina Pasqual, Francesca Tomasi, and Fabio Vitali. 'Expressing Without Asserting in the Arts'. In *Proceedings of the 18th Italian Research Conference on Digital Libraries*, edited by Giorgio Maria Di Nunzio, Beatrice Portelli, Domenico Redavid, and Gianmaria Silvello, 3160:1–9. CEUR, 2022.

[9] Doerr, Martin, Christian-Emil Ore, Pavlos Fafalios, Athina Kritsotaki, Stephen Stead et al. 'Definition of the CRMinf: An Extension of CIDOC-CRM to Support Argumentation', 2023. https://cidoc-crm.org/crminf/

[10] Drucker, Johanna. 'Humanities Approaches to Graphical Display'. *Digital Humanities Quarterly* 5, no. 1 (2011).

[11] Gius, Evelyn, and Janina Jacke. 'The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis'. *International Journal of Humanities and Arts Computing* 11, no. 2 (2017): 233–254.

[12] Hastings, Janna, and Stefan Schulz. 'Representing Literary Characters and Their Attributes in an Ontology'. In *Proceedings of the Joint Ontology Workshops*, edited by Adrien Barton et al., 1–10. CEUR 2518, 2019.

[13] Jannidis, Fotis. 'Character'. In *The Living Handbook of Narratology*, edited by Peter Hühn. Hamburg: Hamburg University, 2013.

[14] Kroon, Fred, and Alberto Voltolini. 'Fictional Entities'. Edited by Edward N. Zalta and Uri Nodelman. *The Stanford Encyclopedia of Philosophy (Fall 2023 Edition)*, n.d. https://plato.stanford.edu/archives/fall2023/entries/fictional-entities/

[15] Masolo, Claudio, Alessander Botti Benevides, and Daniele Porello. 'The Interplay between Models and Observations'. *Applied Ontology* 13, no. 1 (2018): 41–71.

[16] Morabito, Raffaele. *Le virtù di Griselda: storia di una storia*. Firenze: Olschki, 2017.

[17] Picchio Simonelli, Maria. *Beatrice nell'opera di Dante e nella memoria europea (1290-1990)*. Firenze: Cadmo, 1994.

[18] Renzi, Lorenzo. *Le conseguenze di un bacio: l'episodio di Francesca nella «Commedia» di Dante*. Bologna: Il Mulino, 2007.

[19] Rockwell, Geoffrey, and Stéfan Sinclair. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press, 2022.

[20] Sanfilippo, Emilio, Antonio Sotgiu, Gaia Tomazzoli, Claudio Masolo, Daniele Porello, and Roberta Ferrario. 'Ontological Modeling of Scholarly Statements: A Case Study in Literary Criticism'. In *Formal Ontology in Information Systems. Proceedings of FOIS 2023*, edited by Nathalie Aussenac-Gilles et al., 349–363. Amsterdam: IOS Press, 2023.

[21] Santagata, Marco. *Boccaccio indiscreto: il mito di Fiammetta*. Bologna: Il Mulino, 2019.

[22] Sartini, Bruno, Sofia Baroncini, Van Erp Marieke, Francesca Tomasi, and Aldo Gangemi. 'Icon: An Ontology for Comprehensive Artistic Interpretations'. *ACM Journal on Computing and Cultural Heritage* 16, no. 3 (2023): 1–38.

[23] Zöllner-Weber, Amélie. 'Formale Repräsentation Und Beschreibung von Literarischen Figuren'. In *Jahrbuch Für Computerphilologie 7*, edited by Georg Braungart et al. Paderborn: Mentis Verlag, 2005.

Representing texts as LOD: a Systematic Literature Review

Michela Bandini[1], Valeria Quochi[2]
[1]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - michela.bandini@ilc.cnr.it
[2]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - valeria.quochi@ilc.cnr.it

ABSTRACT

Despite the growing interest in publishing linguistic data as Linked Open Data, the publishing of ancient language corpora for the Semantic Web is still challenging. This contribution describes a systematic literature review on the representation of corpus data as Linguistic Linked Open Data, focusing especially on models and (data) granularity. Our goal is to gain insights into the advantages and disadvantages of the different approaches. Here we present our systematic review methodology and some initial results.

KEYWORDS

Linked Open Data; corpora; ancient languages; systematic literature review.

1. INTRODUCTION

A trend has gained increasing attention in the representation and publication of language datasets as Linked Open Data (LOD), primarily for Knowledge Representation (KR) and Natural Language Processing (NLP) purposes. In recent times, LOD and Semantic Web (SW) technologies have also captured the interest of digital humanists, with an expanding variety of data sources being published as LOD. The vast majority of linguistic resources in the LOD cloud [5] are dictionaries, lexica, thesauri, terminologies, and (controlled) vocabularies. There is a recent growing interest in publishing text corpora as Linguistic LOD (LLOD) both in NLP and in Digital Humanities (DH) communities. Linked data, in fact, "allows better data integration than existing models of linguistic data, due to the ecosystem of tools provided by the Semantic Web" and "enable[s] better resource interoperability" [11: 315]. As part of an ongoing project on the digital representation of scholarly knowledge about archaic languages and cultures, one of the main goals of the present contribution is to explore the possibilities for representing and publishing corpora of ancient inscriptions (mostly available in `XML` TEI-Epidoc or in database formats) as LLOD. We thus started with a systematic literature review on this topic. The review's methodology follows 3 generic steps, described in detail in the rest of the paper.

2. REVIEWING METHODOLOGY

Defining terms and questions of the research

This systematic review addresses works and projects in which text corpora are represented as Linked Open Data. We, therefore, developed this review to understand:

what are the most relevant works that have already attempted to transform - or represent - (text) corpora as LOD;
what are the models and formats already in use by the digital humanities community to represent corpora as LOD;
how can we classify projects according to the model used for representation, to the granularity of the representation and to the purpose (or research sub-community).

Given the high number of studies and articles regarding the publication of data following LOD principles, we decided to adopt a systematic approach to help us focus on relevant studies. We searched specific digital libraries, used seed keywords and authors, and applied several criteria to filter the results to concentrate on the most relevant for our purposes. Figure 1 shows the workflow which explains the steps followed for the selection of works used in this review.

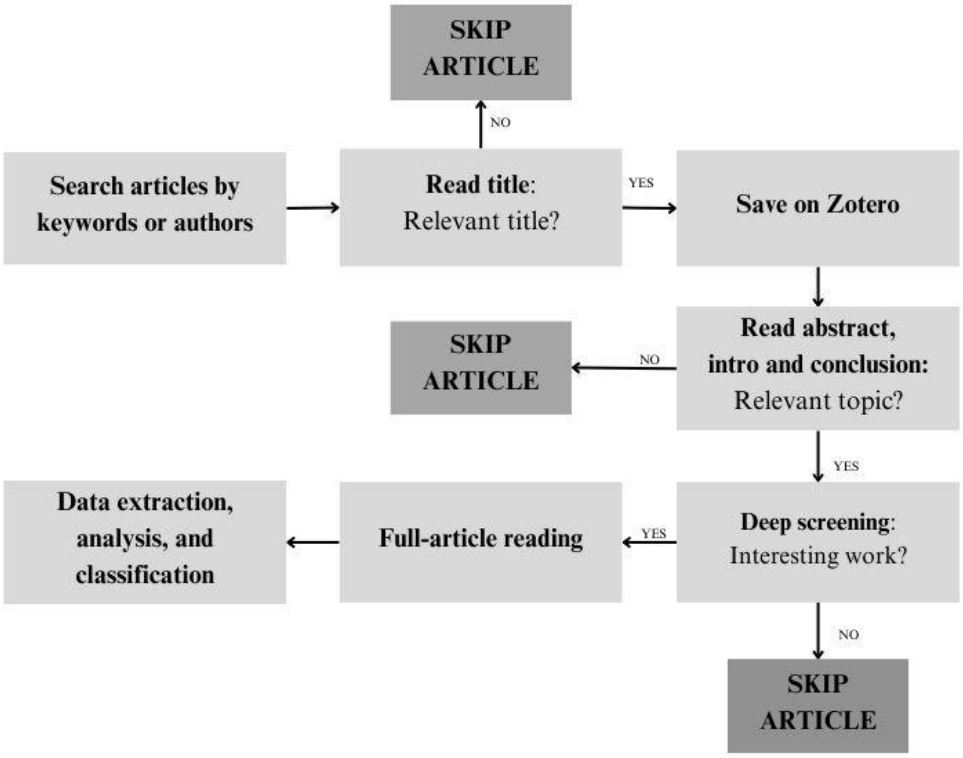


Figure 1. Reviewing workflow

Defining criteria: sources, keywords/seed terms and authors

We mainly applied regular and advanced online searches on digital libraries such as *ACL Anthology, DBPL, Google Scholar, IEEE Xplore, and Semantic Scholar*. We used 3 different seed keywords combined with other terms as reported in Table 1. We opted for 2 groups of terms to be inclusive and relevant. Terms in the first group were considered seed keywords (second column), main and generic words concerning core or general topics such as "Linked Open Data", "LOD", or "XML/RDF"; in the second group terms were related to more specific topics strictly linked with our project such as "ancient languages" or "corpora". The idea was to use a seed keyword combined with a more detailed term to gather projects as relevant as possible to our interests. We, indeed, tried to broaden the research with a multi-term search using words closely related to the main focus of our project. Other extra keywords, such as "POWLA" or "NIF", as well as seed authors[1] were used for targeting known formats/models directly. The search conducted on digital libraries was also filtered by date from 2000 up to today and by language (we only opted for English and Italian works). The results were sorted by relevance[2] and/or by the number of citations (when possible). Whenever the number of results pages was high, we only focused on the first 15 pages per search.

Linked Open Data		**LOD**		**XML/RDF**		
ancient languages	linked open data	ancient languages	LOD	from	XML	RDF
corpora	linked open data	corpora	LOD	transitioning	XML	RDF
edition	linked open data	edition	LOD	convert	XML	RDF
historical text	linked open data	historical text	LOD	corpora		RDF
transform	linked open data	transform	LOD	conversion		RDF

Table 1. Seed Keywords combinations

[1] These are authors strictly related to LOD works on corpora and are the most cited authors in articles about this topic.

[2] It refers to the plug-in functionality of the respective digital libraries where it's possible to reorder search results by relevance or pertinence.

3. LITERATURE SCREENING AND ASSESSMENT

Screening for Inclusion

Title Reading. We focused on conference papers, journal articles, extended abstracts, dissertations, specific case studies, and book chapters. For each search, we read through all the resulting titles and ignored all those papers which clearly were not relevant (e.g. generic and theoretical papers related to topics such as what is LOD, the RDF format, etc.). All other articles, 219, were passed to the next stage and saved in a dedicated Zotero library[3].

Abstract, introduction, and conclusion reading. We proceeded with the screening of the abstract, introduction, and conclusion of the items saved in the Zotero library, to decide which ones were truly relevant for our research. At this stage, we only included works that described some model or approach to represent or convert (possibly annotated) texts in compliance with L(O)D principles, i.e. papers that even generically described the representation of textual data for the Semantic Web. At the end of this step, we were left with 136 relevant papers, and excluded 83 papers which:

- illustrated general and theoretical topics (e.g. description of formats such as XML or RDF, etc.). Yet, we took into account theoretical papers regarding corpora (e.g. state-of-art about POWLA [2] or Ligt [4]);
- provided not interesting works or not relevant topics (e.g. describing the process used in the digitization of data, not related to LOD);
- provided a LLOD-compliant model or activity clearly not related to texts (e.g. OntoLex Lemon, SKOS, etc.[4]).

Deep reading. The remaining papers were skimmed through entirely to determine whether they were actually relevant. This time we also focused on specific keywords present inside the texts such as “XML”, “text”, “corpora” or “corpus”, “sentences”, “books”, “manuscripts”, etc. We excluded another 59 articles which:

did not target text-corpora, but rather corpus-derived data represented as lexicons, or CSV / TSV data.
mentioned textual data, but in fact dealt with platforms, website implementation, or specific ontologies.

Quality and Eligibility Assessment

Full-article reading. Following the second screening, 77 articles remained for deep full-text reading aimed at further assessing the quality and eligibility of the works and eventually excluded a few other non-relevant works. 12 relevant works were theoretical papers on LOD representation models for texts. The rest focused on case studies or project-related works. The assessment and analysis were performed independently by the authors of this paper, and disagreement was resolved through discussion.

4. DATA ANALYSIS

The 77 remaining papers were analyzed and categorized based on two key criteria: the level of granularity in data representation and the models and formats used for representing texts within the Semantic Web. Our primary goal was to elucidate prevalent practices and identify trends in making annotated texts available on the Semantic Web. The discussion below synthesizes the most significant findings from our analysis.

Granularity of the data representation. From the perspective of granularity representation of data, many surveyed papers represent datasets at a document level, i.e. as bibliographic entities or cultural objects, without detailing the representation of the textual data thereby contained, i.e. sequences of linguistic signs. For example, in the *Mapping Manuscript Migrations (MMM) project* [9] manuscripts are represented as bibliographic entries, i.e. at metadata level (e.g. author, production place, production data, language, etc.).

Most analyzed papers feature a “partial” representation of text contents as LOD, that is: only some predetermined extracted text parts are represented as `RDF` triples, such as named entities or events, which are then linked to some external KB/KG. For example, in [1] tokens are extracted from Arabic sentences and automatically mapped to DBPedia for generating semantic triples as enrichments of the original text, with text documents and triple datasets remaining distinct.

[3] The exported and versioned dataset of the Zotero library discussed in this paper is available on Zenodo for reference and reproducibility purposes. See https://zenodo.org/doi/10.5281/zenodo.10978178. The bibliographic entries within the library are categorized into “relevant” and “not-relevant” works. All tags used for classification and analysis have been preserved in the exported dataset.

[4] Yet, some papers were retained if they discussed projects or platforms that integrate various types of data, including text corpora; for instance the *LiLa: Linking Latin* project [13].

In a few other projects, mainly those related to NIF and CoNLL-RDF, the representation is more granular. The *Machine Translation and Automated Analysis of Cuneiform Languages project (MTAAC project)* [6], for instance, employs the CoNLL-RDF [3] model to represent texts as LOD. In this context, the work is an example of the depth of the representation we are interested in, which includes sentence offsets, tokens, morphological information and word order, as shown in Figure 2, extracted from an actual example provided in [6: 2441][5]. Within this LOD representation, the authors can represent the cuneiform corpus providing details about the beginning and end of sentences, their components and words' morphological information, and even the word's order thanks to specific CoNLL-RDF attributes. To interpret the code provided in the figure below, the sentence is defined with nif:Sentence, word is defined as a nif:Word, followed by its conll:WORD, other annotations in alphabetical order of their properties are provide, concluding with a nif:next statement pointing to the next word in the sentence. The relationship between words and sentences is established by conll:HEAD and conll:WORD. The attribute nif:nextSentence is used in case there are more sentences following the one represented.

```
@prefix : <http://oracc.museum.upenn.edu/etcsri/Q000935#> .
@prefix conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#> .
@prefix rdf: <http://www.w3.Org/1999/02/22-rdf-syntax-ns#> .
@prefix terms: <http://purl.org/acoli/open-ie/> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix rdfs: <http://www.w3.Org/2000/01/rdf-schema#> .

:s2_0 nif:nextSentence :s3_0 .

:s3_0 a nif:Sentence .
:s3_1 a nif:Word; conll:WORD "lu₂";
terms:lemma <http://psd.museum.upenn.edu/epsd/epsd/e3356>; conll:BASE "lu₂";
conll:CF "lu"; conll:EPOS "n"; conll:FORM "lu₂";
conll:GW "person";
conll:HEAD :s3_0; conll:ID "1"; conll:LANG "sux"; conll:MORPH "N1=lu";
conll:MORPH2 "N1=stem"; conll:NORM "lu"; conll:POS "N"; conll:SENSE "person";
nif:nextWord :s3_2 .

:s3_2 a nif:Word; conll:WORD "e₂";
terms:lemma <http://psd.museum.upenn.edu/epsd/epsd/ell66>; conll:BASE "e₂ "; conll:CF "e"; conll:EPOS "n"; conll:FORM "e₂";
conll:GW "house";
conll:HEAD :s3_0; conll:ID "2"; conll:LANG "sux"; conll:MORPH "N1=e"; conll:MORPH2 "N1=STEM"; conll:NORM "e"; conll:POS "N";
conll:SENSE "house, temple";
nif:nextWord :s3_3 .

:s3_3 a nif:Word; conll:WORD "{d}nanna"; conll:BASE "{d}nanna";
conll:CF "Nanna"; conll:EPOS "DN"; conll:FORM "{d}nanna\\gen\\abs";
conll:GW "1";
conll:HEAD :s3_0; conll:ID "3"; conll:LANG "sux"; conll:MORPH "N1=Nanna.N5=ak.N5=Ø";
conll:MORPH2 "N1=name.N5=gen.N5=abs"; conll:NORM "Nanna.ak.Ø"; conll:POS "DN"; conll:SENSE "1" .
```

Figure 2. Example of CoNLL-RDF representation of textual corpora representation in MTAAC project

Lastly, the representation of linguistic corpora according to POWLA [2] generally also extends from sentence to token level and can include linking to external resources to provide richer morphological and linguistic information. As shown in Figure 3, in the LASLA corpus of the *LiLa: Linking Latin* project [7][6], the document has different layers of representation to encode sentences and tokens. Specific properties are used to specify detailed information about the structure of the text, such as the beginning or the ending of the sentences and their token order.

Models and Formats. Regarding this criterion, a number of surveyed works can be considered precursors to LOD representation either because they predate the definition of LLOD or because they rely on customizations of the XML formats, allowing direct use of RDF within TEI documents by exploiting **RDFa**. In such cases, RDF triples are directly encoded inline in the XML documents. For instance, the *Diachronic Spanish Sonnet Corpus (DISCO)* [16] makes use of TEI/XML for the digital edition of poems by Spanish and Latin American authors from the 15th to the 19th century and includes RDFa attributes to incorporate links to external metadata sources, such as VIAF and Wikidata for author biographical information (e.g. birthplace, date of birth and death, profession). Figure 4 below displays a simplification of

[5] This code-snippet is a re-elaboration of the provided in Figure 1 of [6: 2441].
[6] This code-snipped is extracted from the "Catullus Catullis" text of the LASLA corpus represented in POWLA (lines 14-26; 39-48; 55-58). See the project's github repository for the full text code: http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus

the original XML representation used for the encoding of the *DISCO*[7] corpus. As we can see, an RDFa layer is encoded with different attributes: with the @typeOf attribute the domain of the properties is declared, with @property the predicates of the RDF triple are defined, @about is used to represent the subject, while its IRI is added with @resource.

```
<http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus> a powla:Document;
  dc:title "Catullus";
  <http://purl.org/dc/terms/creator> <http://www.wikidata.org/entity/Q163079> .

<http://lila-erc.eu/data/corpora/Lasla/id/corpus> powla:hasSubDocument <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus> .
<http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer> a
    lila_corpus:CitationStructure;
  dc:title "Catullus_Catullus Sentence Layer";
  dc:description "Catullus_Catullus Sentence Layer";
  powla:hasDocument <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus>;
  lila_corpus:first <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_1>;
  lila_corpus:isLayer <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_1>,
  [...]
    <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_14> .

    <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_1>
  a lila_corpus:citationUnit;
  rdfs:label "Sentence 1";
  lila_corpus:hasRefType "Sentence";
  lila_corpus:hasCitLevel "1"^^xsd:int;
  lila_corpus:hasRefValue "Sentence_1";
  powla:hasChild <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000001>,
    <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000002>,
    [...]
    <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000009>;
  lila_corpus:first <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000001>;
  lila_corpus:last <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000009>;
  powla:next <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_2> .
```

Figure 3. Example of POWLA representation of LASLA corpus in LiLa project

```
<person xml:id="disco_100n" about="disco:100n" typeof="foaf:Person">
    <idno cert="high"
        property="rdfs:seeAlso"
        resource="https://viaf.org/viaf/29108480"/>
    <persName type="full">
        <forename property="foaf:givenName">Antonia</forename>
        <surname property="foaf:familyName">Díaz de Lamarque</surname>
    </persName>
    <sex property="foaf:gender" content="F"/>
    <birth>
        <location>
            <placeName>
                <settlement property="schema:birthPlace">Marchena (Sevilla)</settlement>
            </placeName>
        </location>
        <date property="schema:birthDate" content="1837" cert="high"/>
    </birth>
    <death>
        <date property="schema:deathDate" content="1892" cert="high"/>
    </death>
    <listBibl rel="blterms:hasCreated">
        <bibl resource="disco:s100n_0335" typeof="schema:CreativeWork">
            <title property="dc:title">A Dios en la Eucaristía</title>
            <title type="incipit" property="dc:alternative">Tu infinito poder en la armonía</title>
        </bibl>
    </listBibl>
</person>
```

Figure 4. Example of TEI/XML-RDFa representation of bibliographical information in DISCO project

Other projects explicitly rely on domain-specific RDF models and/or vocabularies, such as **CoNLL-RDF** [3], used to represent linguistically annotated natural languages. The *MTAAC project* [6] is a nice example of the application of this model to represent linguistically annotated text as LOD. According to this first and broad analysis, this model can be considered one of the most convenient solutions for representing textual data or sentences, given the detailed possibilities in granularity representation.

[7] This code-snipped is extracted from the DISCO project's GitHub public repository (READ-ME section). See https://github.com/pruizf/disco/tree/v2.1

CoNLL-RDF was developed based on the **NLP Interchange Format (NIF)** [8], a stand-off representation model for the integration of corpus data into the Semantic Web, especially devised to leverage NLP tools over L(O)D. [17], for example, successfully employs it to convert and publish the "Manually Annotated Subcorpus (MASC) of the American National Corpus" at a good granular level, as stated previously. However, it looks like this format is not very popular: we found only one other work that exploits it [15]. As also stated in [10], there seems to be a sort of dispreference of NIF over other solutions such as CoNLL-RDF, apparently because it does not provide sufficient support for the annotation of morphological traits and for the internal structure of words. In detail, NIF is an RDF-based format for describing strings in text documents, and its classes and properties are defined in the NIF Core Ontology. **Figure 5** is a code excerpt extracted from the NIF edition of the Brown corpus published in 2015 [10][8]. As shown in the figure, the context, which usually describes the whole document's text, is created with nif:Context. Then each node is associated with a nif:String to represent a textual chunk, in our example: sentence (nif:Sentence) and words (nif:Word). For each textual chunk, properties are assigned to provide information about its beginning and the ending, and the representation of the actual string.

Several papers in our review, instead, represent linguistic corpora according to **POWLA,** an OWL2/DL vocabulary for linguistic annotations based on the LAF ISO standard, made to support any kind of text-oriented annotation [2]. It is exploited in many projects throughout the digital humanity community; we count at least 10 papers in our screening, 7 of which however are about the *LiLa: Linking Latin* ERC project, which seeks to interlink and publish in a machine-actionable way different Latin language data resources [12].

Other relevant projects do not adhere to any of the previously mentioned models and describe other, mostly custom or proprietary, models and formats. For instance, the *Poetry Standardization and Linked Open Data* project develops a specific ontology for the semantic representation of European poetry [14]. Other works, like the project *Orlando: Women's Writing in the British Isles Project* [18], provide some solutions to partially convert XML documents to RDF.

```
<http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_161>
        a nif:String , nif:Context , nif:OffsetBasedString ;
        nif:isString """The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary
        election produced ``no evidence'' that any irregularities took place. [...]"""^^xsd:string ;
        nif:beginIndex "0"^^xsd:int ;
        nif:endIndex "161"^^xsd:int ;
        nif:sourceUrl <http://icame.uib.no/brown/bcm.html>

<http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_155>
        a nif:String , nif:Sentence , nif:OffsetBasedString ;
        nif:anchorOf """The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary
        election produced ``no evidence'' that any irregularities took place."""^^xsd:string ;
        nif:referenceContext <http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_161> ;
        nif:beginIndex "0"^^xsd:int ;
        nif:endIndex "155"^^xsd:int .

<http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_3>
        a nif:String , nif:Word , nif:OffsetBasedString ;
        nif:anchorOf "The"^^xsd:string ;
        nif:referenceContext <http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_161> ;
        nif:oliaLink brown:AT ;
        nif:nextWord <http://brown.nlp2rdf.org/corpus/a01.xml#offset_4_10> ;
        nif:sentence <http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_155> ;
        nif:beginIndex "0"^^xsd:int ;
        nif:endIndex "3"^^xsd:int .
```

Figure 5. Example of NIF representation of the Brown corpus

5. CONCLUSION

This contribution presented the initial results of a systematic literature review on models and representation formats of linguistic textual data as LLOD. Although the interest in the task seems to be growing, in practice the actual adoption of these practices remains limited, with few projects publishing linguistic text corpora as LOD. This raises concerns about the reasons and their real-world viability or utility.

Our exploration of models and formats for LOD representation for linguistic corpora reveals a restricted range of existing approaches. Three models, in particular, stand out as the most interesting and used: CoNLL-RDF and NIF, which seem to demonstrate their effectiveness in representing linguistically annotated corpora, especially within NLP; and the POWLA ontology/model, which is a preferred choice in digital humanities projects, notably in the *LiLa Knowledge Base*.

[8] The snippet in Figure 5 is taken from https://bpmlod.github.io/report/nif-corpus/index.html.

Looking ahead, we plan to expand our review to include not only published papers but also datasets themselves, potentially enriching our understanding of the field. Many datasets, as a matter of fact, may not be described in publications, but could instead be available in discipline-specific or institutional data repositories, such as Zenodo and Research Infrastructure repositories. This broader approach may provide a more comprehensive picture of how linguistic data is represented and utilized in current research.

6. ACKNOWLEDGEMENTS

This work is carried out in the context of the PRIN 2017 "Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models" (no. 2017XJLE8J) funded by the Italian Ministry of University and Research. The DigItAnt platform is also supported by CLARIN-IT.

REFERENCES

[1] Bouziane, Abdelghani, Bouchiha Djelloul, and Doumi Noureddine. '"Annotating Arabic Texts with Linked Data."' 2020 4th International Symposium on Informatics and Its Applications (ISIA), 2020, 1–5.

[2] Chiarcos, Christian. 'POWLA: Modeling Linguistic Corpora in OWL/DL.' In *The Semantic Web: Research and Applications*, edited by Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, 7295:225–239. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012.

[3] Chiarcos, Christian, and Luis Glaser. '"A Tree Extension for CoNLL-RDF."*' In Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7161–7169. European Language Resources Association, 2020.

[4] Chiarcos, Christian, and Maxim Ionov. 'Ligt: An LLOD-Native Vocabulary for Representing Interlinear Glossed Text as RDF', 2019.

[5] Chiarcos, Christian, Bettina Klimek, Christian Fäth, Thierry Declerck, and John Philip McCrae. 'On the Linguistic Linked Open Data Infrastructure'. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 8–15. Marseille, France: European Language Resources Association, 2020.

[6] Chiarcos, Christian, Émilie Pagé-Perron, Ilya Khait, Niko Schenk, and Lucas Reckling. 'Towards a Linked Open Data Edition of Sumerian Corpora'. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

[7] Fantoli, Margherita, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 'Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin'. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, 26–34. Marseille, France: European Language Resources Association, 2022.

[8] Hellmann, S., J. Lehmann, S. Auer, and M. Brümmer. 'Integrating NLP Using Linked Data'. In The *Semantic Web – ISWC*, Vol. 8219. Berlin, Heidelberg: Springer, 2013.

[9] Hyvönen, Eero, Esko Ikkala, Mikko Koho, Jouni Tuominen, Toby Burrows, Lynn Ransom, and Hanno Wijsman. 'Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research'. In *The Semantic Web – ISWC*, edited by Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani, 12922:615–630. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021.

[10] Khan, Fahad Anas, Christian Chiarcos, Thierry Declerck, Daniela Gîfu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, et al. 'When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Data'. *Semantic Web* 13 (2022): 1–64.

[11] McCrae, John P., Steven Moran, Sebastian Hellmann, and M. Brümmer. 'Multilingual Linked Data'. *SemanticWeb* 6 (2015): 315–317.

[12] Passarotti, Marco, Elena Litta, Flavio Massimiliano Cecchini, Matteo Pellegrini, Giovanni Moretti, Paolo Ruffolo, and Giulia Pedonese. 'The LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. Architecture and Current State'. In *Elsevier Guest Seminar Series*, 2022.

[13] Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 'Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin'. *Studi e Saggi Linguistici* 58, no. 1 (2020): 177–212. https://doi.org/10.4454/ssl.v58i1.277

[14] Platas, María Luisa Diez, Salvador Ros, Elena González-Blanco, Helena Bermúdez, and Oscar Corcho. 'The POSTDATA Network of Ontologies for European Poetry', 2019.

[15] Rezk, Martín, Jungyeul Park, Yoon Yongun, Kyungtae Lim, John Larsen, Young Gyun Hahm, and Key-Sun Choi. 'Korean Linked Data on the Web: Text to RDF'. In *Semantic Technology*, edited by Hideaki Takeda, Yuzhong Qu, Riichiro Mizoguchi, and Yoshinobu Kitamura, 7774:368–374. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[16] Ruiz, Fabo Pablo, Helena Bermúdez Sabel, Clara Martínez Cantón, e Elena González-Blanco. 'The Diachronic Spanish Sonnet Corpus: TEI and Linked Open Data Encoding, Data Distribution, and Metrical Findings'. *Digital Scholarship in the Humanities* 26, no. Supplement 1 (2021): i68-i80.

[17] Siemoneit, Benjamin, John Philip McCrae, and Philipp Cimiano. ‘Linking Four Heterogeneous Language Resources as Linked Data’. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, 59–63. Beijing, China: Association for Computational Linguistics, 2015.

[18] Simpson, John, and Susan Brown. ‘From XML to RDF in the Orlando Project’. In *2013 International Conference on Culture and Computing*, 194–195, 2013.

Structuring Authenticity Assessments on Historical Documents using LLMs

Andrea Schimmenti[1], Valentina Pasqual[2], Francesca Tomasi[3], Fabio Vitali[4], Marieke van Erp[5]
[1] University of Bologna, Italy - andrea.schimmenti2@unibo.it
[2]Digital Humanities Advanced Research Center, (/DH.arc), FICLIT, University of Bologna, Italy - valentina.pasqual2@unibo.it
[3]University of Bologna, Italy - francesca.tomasi@unibo.it
[4]University of Bologna, Italy - fabio.vitali@unibo.it
[5]DHLab - KNAW Humanities Cluster, Netherlands - marieke.van.erp@dh.huc.knaw.nl

ABSTRACT

Given the wide use of forgery throughout history, scholars have and are continuously engaged in assessing the authenticity of historical documents. However, online catalogues merely offer descriptive metadata for these documents, relegating discussions about their authenticity to free-text formats, making it difficult to study these assessments at scale. This study explores the generation of structured data about documents' authenticity assessment from natural language texts. Our pipeline exploits Large Language Models (LLMs) to select, extract and classify relevant claims about the topic without the need for training, and Semantic Web technologies to structure and type-validate the LLM's results. The final output is a catalogue of documents whose authenticity has been debated, along with scholars' opinions on their authenticity. This process can serve as a valuable resource for integration into catalogues, allowing room for more intricate queries and analyses on the evolution of these debates over centuries.

KEYWORDS

LLM; Semantic Web: critical debate; forgery; knowledge extraction.

1. INTRODUCTION

Historical authenticity assessment is a scholarly practice used to determine the authenticity of historical documents. Scholars from different humanities and scientific disciplines (e.g. Diplomatics, Palaeography, Philology, History, Forensics) have contributed to the task [2]. Various scholars arrive at divergent and possibly contrasting conclusions due to considering different evidence. Inherent factors contributing to this diversity include historical uncertainty, gaps in documentary transmission, and subjectivity [3, 9]. For example, the *Donation of Constantine* has been studied by several scholars from different perspectives, and its authenticity has been widely discussed for centuries[1]. The debate revolves around this being a supposed decree by Constantine from the 4th century until scholar Lorenzo Valla claimed it was a later forgery at the beginning of the Renaissance. To the best of our knowledge, such complex information is not included in digital libraries, catalogues and archives. Considering the provided guiding example, the document can be found in Wikidata and DBpedia, but no mention of the debate is available in a structured format. Considering the intricate language and data sparseness of humanities discourse, the recent spread of Large Language Models (LLMs) widens the possibility of retrieval from unstructured natural language descriptions. Many tests have proved that LLMs perform promisingly when dealing with structured data extraction from unstructured text [13, 14, 16]. This work tackles the following research question: can LLMs be useful inside a pipeline as tools for extracting and classifying authenticity assessment debates, with the ultimate goal of structuring such claims into a KG? Our approach to answering this question revolves around analysing the authenticity assessments of Wikipedia articles using two different LLMs. We used the outputs of the LLMs to generate a KG. For the Donation of Constantine case study, we show how a structured authenticity assessment KG can automatically map a forgery debate.

2. RELATED WORK

Knowledge Graph (KG) generation from text has been the aim of many researchers inside and outside the Semantic Web community. For example, REBEL, trained on multiple languages (using a set of Wikidata properties and classes as ontology), is an example of state-of-the-art performance in generating simple KGs [12]. Since 2017, many NLP tasks have seen general improvements after introducing the Transformers architecture and pre-trained models such as BERT [8]. With their massive training on multiple corpora, Large Language Models can be useful, especially in cases where large datasets are unavailable for downstream tasks, e.g. KG generation. LLMs have also been tested for KG extraction from unstructured

[1] https://en.wikipedia.org/wiki/Donation_of_Constantine

text, especially for KG generation and enrichment. BERT proved to have good generalisation capabilities, even with a few examples in the training data using in-context learning (i.e. describing a task alongside one or multiple examples) [14]. LLMs, such as the GPT family, have been tested for multiple tasks, achieving state of art results with Zero or Few-Shot learning [4]. However, LLMs are not completely reliable as a source of factual knowledge for sturdy KB enrichment, relying mostly on superficial natural language information. Yet, they perform well compared to state-of-the-art results in many NLP tasks [16]. Text2KG, a benchmark for evaluating LLM results over generating KGs given a simplified ontology and a few examples, has shown promising results when evaluated on two different datasets, Wikidata-TekGen and DBpedia-WebNLG [15]. The prompts used a simplified version of the Wikidata or the DBpedia set of properties, where a property was described as, e.g. "creatorOf(person, book)", a set of terms for the possible classes and a set of examples [15]. This type of in-context learning also shows interesting results when dealing with KG generation using GPT-3 [14]. In general, the application of LLMs in Knowledge Engineering is being discussed by the community with multiple open questions on how they can and should be used and how reliable they can be when dealing with knowledge engineering in general [1].

3. APPROACH

From a selection of Wikipedia articles regarding documentary forgery and medieval charters[2], we extracted the document description and those sections tackling authenticity assessment. The test involved three tasks whose output was promptly evaluated. The output data created a KG using *forgont* as an ontology. For these tasks, we chose to test two different LLMs: GPT-4[3] and Llama version 2 70b[4]. The workflow of the experiment is as follows:

Data modelling. An ontology, called *forgont,* was defined based on a selection of scholarly articles [6,11], a catalogue describing 153 known forgeries from Styria [10], and collaboration with an expert Diplomatist. The ontology *forgont* represents authenticity assessment claims using the Named Graphs [5] structure as a reification method to represent both each (possibly concurrent) claim content and its contextual information [7]. The content of the claim addresses the following basic information about the document in the scholar's opinion: document authenticity classification (e.g. authentic, forgery, suspicious forgery), date and place of creation, author, and reason of the document. Additionally, contextual information about the claim is recorded, specifically the author of the claim, the document features observed by the scholar (e.g. ink, paper, style), the evidence found (e.g. inconsistent contents with other documents, lack of genuine witnesses) to reach a certain conclusion (e.g. the document is a forgery), and relevant sources[5]. The *forgont* structure served as a schema to define the prompts;

LLM-assisted data extraction.

Task 1: Document metadata extraction. The LLM prompt asked to identify and extract a distinct set of properties from the text, namely the "document title", "document type", (alleged) "creation date", (alleged) "creation place," and (alleged) "author". The model is required to extract the metadata as presented *within* the forged document: in the case of the *Donation of Constantine*, the document was allegedly created during Constantine's reign, while it was in reality produced much later;

Task 2: Named Entity Recognition and Claim Identification. The model is tasked to extract from the text every entity that is associated with a claim that discusses the document's authenticity;

Task 3: Claim classification. Given the entity and the claim from the previous task, the model is instructed to classify the claims into three possible categories: "Authentic", "Forgery", and "Suspicious". The output is an object containing the author, class, opinion and source;

Evaluation. LLM's performance for the metadata extraction and claim classification is measured using common metrics: precision, recall and F_1-score, while the claim identification is evaluated by the raw percentage of success. The number of ground truths was manually calculated, corresponding to 296 claims. The evaluation compares the two selected LLMs;

Contents structuring and Knowledge Graph generation. The contents extracted and classified by the winning LLM Data have been saved in JSON format and converted into RDF via Python scripts[6]. In the process, data have been cleaned and type-validated. In particular, people were semiautomatically reconciled against authority files (e.g.

[2] https://en.wikipedia.org/wiki/Category:Document_forgery;
https://en.wikipedia.org/wiki/Category:Medieval_charters_and_cartularies and their subcategories
[3] https://openai.com/research/gpt-4
[4] https://huggingface.co/meta-llama/Llama-2-70b
[5] The complete documentation of both *forgont* model and relative Knowledge Base can be found at:
https://github.com/ValentinaPasqual/forgont/.
[6] All scripts are stored in the Github folder of the project.

Viaf and Wikidata), dates were converted into machine-readable format, and places were reconciled and validated against Geonames. A human qualitative check of the dataset was performed to ensure quality. All claims were converted into RDF following the *forgont* schema.

4. EVALUATION

First task: metadata extraction. The data extraction task was evaluated on a corpus of 57 entries, assessing five key metadata properties: Title, Type, Date, Place, and Creator. Each entry of the corpus[7] contains the internal ID of the document, its Wikipedia link, its textual description and a selection of the sections describing the authenticity assessment debate. The criteria alignment is the following:

- Positive: The value was present and correctly identified in the text.
- Negative: The value was absent and correctly reported as not mentioned in the text.
- False Positive: An incorrect value is reported as present despite being absent in the text.
- False Negative: The value was present in the text but failed to be identified.

We computed Precision, Recall and F-1 score (see Table 1).

First task	**Title**		**Type**		**Date**		**Place**		**Creator**	
	GPT-4	**Llama-2**	**GPT-4**	**Llama-2**	**GPT-4**	**Llama-2**	**GPT-4**	**Llama-2**	**GPT-4**	**Llama-2**
Precision	0.98	0.98	0.94	0.96	0.89	0.78	0.95	0.78	0.79	0.80
Recall	0.99	0.99	0.99	0.99	0.93	0.99	0.92	0.99	0.95	0.99
***F*$_1$-score**	0.99	0.99	0.97	0.98	0.91	0.87	0.95	0.88	0.87	0.89

Table 1: Precision, Recall, F_1 *(Metadata Extraction)*

Titles and **Types** were mostly correctly identified and reported. The **Date** extraction performance implies some challenges in accurately extracting alleged dates, mistaking it for the forgery date. The models' performance vastly differed when identifying the alleged **Place** of origin or reference within the documents. The **Creator's** value was, as the date, sometimes confused with the author of the forgery. This last value gives much space for improvement. A drastic difference in performance is noticeable when authorship was not explicitly mentioned or was ambiguously referenced.

Second task: Named Entity and Claim Identification. This evaluation assessed if a model correctly recognised a named entity claiming something about a document's authenticity. An output is considered faulty if the entity was wrongly identified and/or if the text related to it was not a real claim, as shown in Table 2.

	Correctly identified claims	**Faulty or unidentified claims (%)**
GPT-4	254	42 (14.1%)
Llama-2	226	70 (23.6%)

Table 2. NER and claim extraction evaluation

[7] The selection is available in Github folder of the project.

Third Task: Claim Categorisation Accuracy. The third task involved a manual quantitative evaluation of the remaining claims (excluding faulty ones). This evaluation aimed to verify the correctness of the claim categorisation. To assess the model's performance in categorising claims as "Authentic", "Forgery", or "Suspicious", metrics such as F1-score, precision, and recall were utilised and computed based on a multiclass confusion matrix (see Table 3). These metrics indicate a high level of accuracy in the model's ability to categorise claims, with good performance in identifying "Forgery" and "Suspicious" claims.

Third task	**Authentic**		**Forgery**		**Suspicious**	
	GPT-4	**Llama-2**	**GPT-4**	**Llama-2**	**GPT-4**	**Llama-2**
Precision	0.87	0.74	0.97	0.92	0.94	0.88
Recall	0.83	0.86	0.90	0.80	0.96	0.88
***F*$_1$-score**	0.85	0.80	0.94	0.85	0.95	0.88

Table 3. Precision, Recall, F_1 (Claim Categorisation Accuracy)

5. RESULT

The resulting KG stores 233 claims discussing the authenticity of 57 documents by 223 authors. Figure 1 exemplifies two claims (out of the 13 extracted) regarding the *Donation of Constantine.* Each claim is modelled as a Named Graph. On the one hand, the document claim - the result of extraction task 1 - contains the alleged author (emperor Constantine), the alleged date of creation (4th century, 300-399) and the document type (charter). On the other hand, the result of tasks 2 and 3 is shown in Valla's claim, which categorises the document as a forgery.

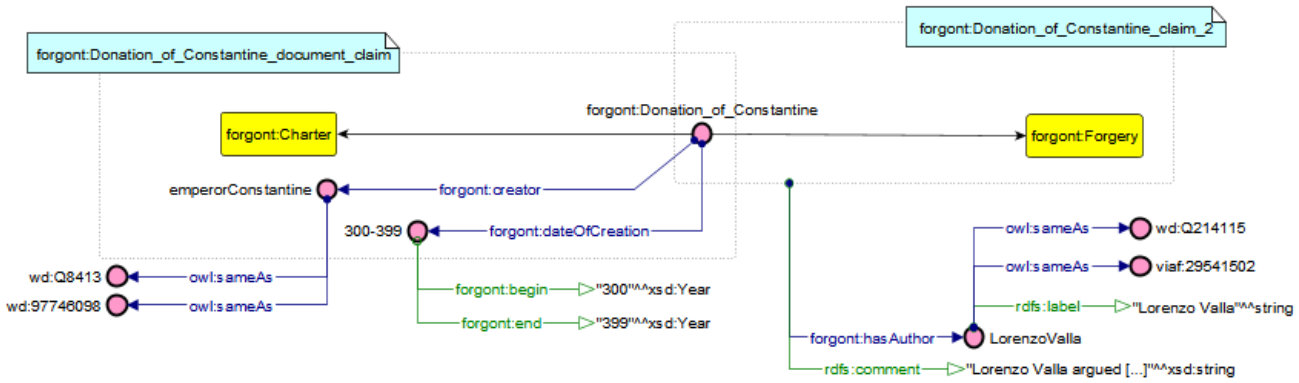


Figure 1. Example of extracted and structured claims regarding the Donation of Constantine

Organising, cleaning, and reconciling data facilitates data retrieval through SPARQL queries and establishes a foundation for in-depth data analysis and visualisation. Extracted claimers have been reconciled to authority files such as VIAF and Wikidata. Life dates of authors have been extracted from Wikidata, and the century of activity has been inferred. Figure 2 illustrates the case of the *Donation of Constantine* and all the claims extracted categories from Wikipedia and categorised by each claimer's century of activity (x-axis) with their opinion (colour). The figure constitutes a possible view of the data extracted and classified by this experiment for this document summarising the debate towards this charter authenticity over the centuries.

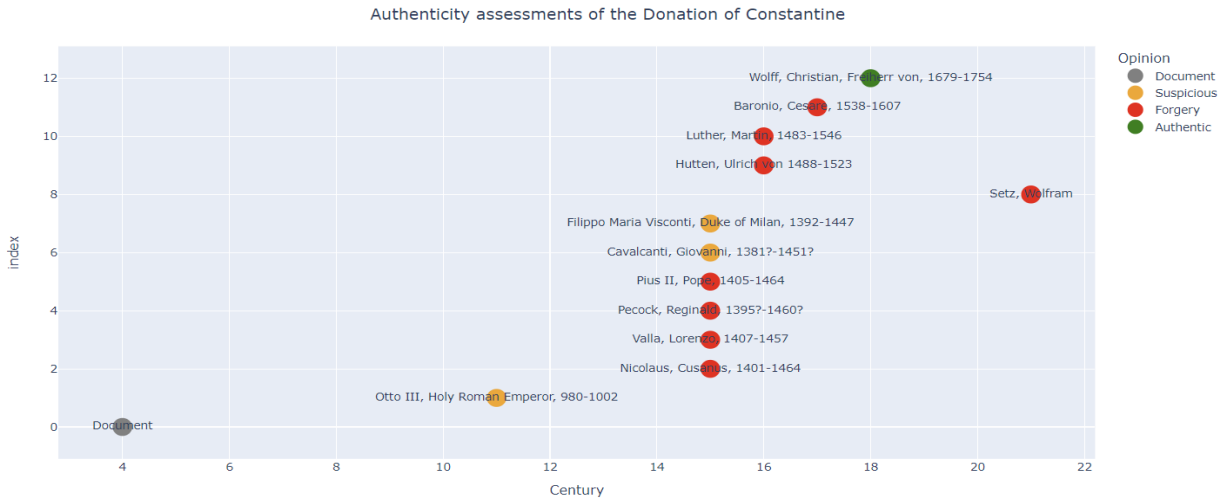


Figure 2. Scatter plot representing claims about the Donation of Constantine

6. CONCLUSION AND FUTURE WORK

This study explored the generation of structured data about documents' authenticity assessment from natural language texts. In particular, the pipeline of this work exploits LLM to select, extract and classify relevant claims, and Semantic Web technologies to structure and validate them. The LLMs offer a simple interface for some of the more complex NLP-related tasks (such as claim extraction and classification) that are otherwise impossible, given the lack of training data for this case study. They are valuable tools inside a pipeline, but using them as an End-to-End model. Errors happen, and avoiding them is impossible without human-in-the-loop approaches. Fine-tuning might be a better solution with a slightly bigger dataset. NLP techniques (e.g. NER) and Semantic Web technologies (e.g. external alignments) are necessary to clean and homogenise the LLM outputs to create a fully structured and interrogable dataset. In conclusion, the *forgont* ontology and the resulting KG demonstrate that highly structured data on the topic of documents' authenticity assessment allows for further queries and analysis on the topic, as shown in Figure 2. This document authenticity assessment serves as a case study highlighting how structuring a critical debate opens a set of new challenges in knowledge extraction. Furthermore, it explores new possibilities in computation, e.g. via SPARQL queries, aiming to understand how scholarly opinions have shaped contemporary knowledge.

7. ACKNOWLEDGEMENTS

This project has been partially funded under the National Recovery and Resilience Plan (NRRP), specifically under Investment I.4.1 - Borse PNRR Patrimonio Culturale. The funding is from the Call for tender No. 351 of 9 April 2022 by the Italian Ministry of Culture, supported by the European Union – NextGenerationEU initiative.
We extend our gratitude to the Università degli Studi di Bologna for their administrative and academic support, particularly in facilitating the course "Patrimonio Culturale nell'Ecosistema Digitale" (Cultural Heritage in the Digital Ecosystem), under the decree Ministeriale n. 351 del 9 aprile 2022.

REFERENCES

[1] Allen, Bradley P., Lise Stork, and Paul Groth. 'Knowledge Engineering Using Large Language Models'. *Graph Data and Knowledge* 1, no. 3 (2023): 3:1-3:19. https://doi.org/10.4230/TGDK.1.1.3

[2] Barone, Nicola. 'Intorno Alla Falsificazione Dei Documenti Ed Alla Critica Di Essi. Memoria Letta All'Accademia Pontaniana Nella Tornata Del 21 gennaio 1912'. In *Atti dell'Accademia Pontaniana*, Vol. 42, 1912. http://www.rmoa.unina.it/4359/

[3] Blau, Adrian. 'Uncertainty and the History of Ideas'. *History and Theory* 50, no. 3 (2011): 358–372. https://doi.org/10.1111/j.1468-2303.2011.00590.x

[4] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 'Language Models Are Few-Shot Learners'. *ArXiv*, July 2020. https://doi.org/10.48550/arXiv.2005.14165

[5] Carroll, Jeremy J., Christian Bizer, Pat Hayes, and Patrick Stickler. 'Named Graphs'. *Journal of Web Semantics* 3, no. 4 (1 December 2005): 247–267. https://doi.org/10.1016/j.websem.2005.09.001

[6] Cau, Ettore. ‘Un falso documento del secolo IX: la donazione di Ottone, Conte Del Seprio, per Il Monasterio di S. Pietro in Ciel d’Oro di Pavia’’. *Istituto Lombardo, Accademia Di Scienze e Lettere. Rendiconti* 122 (1998): 181–196.

[7] Daquino, Marilena, Valentina Pasqual, and Francesca Tomasi. ‘Knowledge Representation of Digital Hermeneutics of Archival and Literary Sources’. *JLIS: Italian Journal of Library, Archives and Information Science = Rivista Italiana Di Biblioteconomia, Archivistica e Scienza dell’informazione: 11, 3, 2020*, no. 3 (2020): 59–76. https://doi.org/10.4403/jlis.it-12642

[8] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. ‘BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding’. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-1423

[9] Gadamer, Hans-Georg. *Truth and Method*. A&C Black, 2013.

[10] Haider, Sigfried. Verzeichnis Der Den Oberösterreichischen Raum Betreffenden Gefälschten, Manipulierten Oder Verdächtigten Mittelalterlichen Urkunden. Oberösterreichisches Landesarchiv, 2022.

[11] Härtel, Reinhard. ‘Il Falso Documento Del Conte Giovanni Di Moggio (875)’. In *Mueç. Societât Filologjiche Furlane/Società Filologica Friulana, XCIV Congrès, Mueç*, edited by Giuliana Pugnetti and Bruno Lucci, 247–252. Udine, 2017.

[12] Huguet Cabot, Pere-Lluís, and Roberto Navigli. ‘REBEL: Relation Extraction By End-to-End Language Generation’. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 2370–2381. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.findings-emnlp.204

[13] Khorashadizadeh, Hanieh, Nandana Mihindukulasooriya, Sanju Tiwari, Jinghua Groppe, and Sven Groppe. ‘Exploring In-Context Learning Capabilities of Foundation Models for Generating Knowledge Graphs from Text’. *ArXiv*, May 2023. https://doi.org/10.48550/arXiv.2305.08804

[14] Meyer, Lars-Peter, Claus Stadler, Johannes Frey, Norman Radtke, Kurt Junghanns, Roy Meissner, Gordian Dziwis, Kirill Bulert, and Michael Martin.. ‘LLM-Assisted Knowledge Graph Engineering: Experiments with ChatGPT’. *ArXiv*, July 2023. https://doi.org/10.48550/arXiv.2307.06917

[15] Mihindukulasooriya, Nandana, Sanju Tiwari, Carlos F. Enguix, and Kusum Lata. ‘Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text’. In *The Semantic Web – ISWC 2023*, edited by Terry R. Payne et al., 247–265. Cham: Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-47243-5_14

[16] Veseli, Blerta, Simon Razniewski, Jan-Christoph Kalo, and Gerhard Weikum. ‘Evaluating the Knowledge Base Completion Potential of GPT’. *ArXiv*, October 2023. https://doi.org/10.48550/arXiv.2310.14771

Valorizzazione di registrazioni storiche di canto lirico nel Web semantico

Marcello Ranieri[1], Angelo Pompilio[2]
[1] Alma Mater Studiorum - Università di Bologna, Italia - marcello.ranieri2@unibo.it
[2] Alma Mater Studiorum - Università di Bologna, Italia - angelo.pompilio@unibo.it

ABSTRACT

Le registrazioni storiche di canto lirico risalenti alla prima metà del Novecento sono conservate prevalentemente su dischi a 78 giri. Ciascuna facciata del disco può contenere un brano della durata massima di circa 4 minuti, pertanto vi sono incisi brani estratti da opere, perlopiù arie, o loro porzioni, con o senza recitativi, spesso con riduzioni o interventi per non eccedere la durata massima consentita. L'etichetta commerciale apposta sul disco riporta di solito come titolo l'incipit testuale del brano selezionato, proposto nella lingua cantata, a volte senza l'indicazione del titolo dell'opera dalla quale è stato estratto. I diversi criteri di selezione dei brani estratti e le diverse lingue di canto fanno sì che lo stesso contenuto musicale possa essere presentato con titoli diversi.

Per la corretta gestione del contenuto di queste registrazioni storiche, i titoli riportati sulle etichette saranno associati ad un'*authority file* di titoli che raggruppi le diverse forme riferibili a un medesimo brano e le riconduca a una sola forma standard.

Identificata l'opera, va indicata la posizione che il brano estratto occupa all'interno dell'opera sulla base di un indice ragionato ottenuto dalla segmentazione del testo musicale completo, realizzata con criteri storicamente fondati.

L'uso dell'*authority file* e dell'indice ragionato richiede un lavoro computazionale che può essere più o meno intenso.

Nell'ecosistema del Web semantico, basato su ontologie formali, il compito è risolto in modo semplice e immediato.

PAROLE CHIAVE

Opera lirica; canto lirico; dischi 78 giri; Linked Open Data; web semantico.

1. INTRODUZIONE

Il canto lirico italiano è stato dichiarato patrimonio immateriale dell'Umanità dall'UNESCO il 6 dicembre 2023[1]. Il riconoscimento del valore universale di questo patrimonio culturale, sviluppatosi nell'arco di circa quattro secoli e profondamente radicato nella vita culturale odierna, conferma la necessità di tutelare e valorizzare il patrimonio documentale che lo testimonia. Conosciamo le registrazioni storiche del repertorio operistico del primo Novecento attraverso le incisioni sonore su dischi, cilindri e nastri magnetici realizzate con la migliore tecnologia disponibile all'epoca. I contenuti di questi supporti non risultano però facilmente accessibili perché richiedono dispositivi di riproduzione oggi difficilmente reperibili sul mercato e costosi. Negli ultimi anni, grazie ad ampie campagne di digitalizzazione, una parte consistente di queste registrazioni storiche risulta accessibile a studiosi ed amatori su collezioni digitali ad accesso pubblico. Da un punto di vista catalografico, è però necessario sviluppare sistemi informativi efficienti ed efficaci, calibrati sulla natura specifica di questi materiali, capaci di consentire una fruizione ottimale di tali contenuti. Si analizza qui il caso delle registrazioni storiche di canto lirico della prima metà del ventesimo secolo, conservate prevalentemente su dischi a 78 giri per minuto. Ciascuna facciata del disco può contenere un brano della durata massima di circa 4 minuti. Di conseguenza sulla facciata del disco sono registrati brani estratti da opere, perlopiù arie, o loro porzioni, con o senza recitativi, spesso con riduzioni o interventi per non eccedere la durata massima consentita. Per la descrizione catalografica dei dischi la fonte informativa primaria prevista dalle norme è l'etichetta apposta[2]. Sull'etichetta commerciale apposta sul disco il titolo riportato è di solito l'incipit testuale del brano selezionato, proposto nella lingua cantata e/o nella traduzione in lingua del paese di edizione, si fa menzione della forma musicale e del titolo dell'opera completa, ma non si indicala posizione che il brano occupa all'interno dell'opera. I diversi criteri con cui i brani sono stati selezionati e le diverse lingue in cui il titolo è presentato fanno sì che un disco possa avere solo in parte lo stesso contenuto di un altro con etichetta simile, oppure che titoli diversi presentino in realtà il medesimo contenuto.

2. IL CONTROLLO DI AUTORITÀ NELLA GESTIONE DELL'INFORMAZIONE

Per la corretta gestione del contenuto di registrazioni storiche, è fondamentale che quanto riportato sulle etichette dei supporti sia soggetto a un controllo di autorità che raggruppi le diverse forme del titolo riferibili a un medesimo brano e le

[1] La notizia è presente sul sito internet del Ministero della Cultura: https://cultura.gov.it/canto-lirico-italiano-patrimonio-unesco.
[2] Cfr. Regole italiane di catalogazione: https://norme.iccu.sbn.it/index.php?title=Reicat/Parte_I/Capitolo_3/3.2/3.2.2 - 3.2.2_D

riconduca a una unica forma standard. Nella catalogazione dei documenti musicali questa funzione è svolta dal titolo dell'opera musicale[3]. Il titolo dell'opera musicale non prevede però l'indicazione della posizione che il brano estratto occupa all'interno dell'opera. Per ottenere questa informazione è necessario esplicitare l'articolazione formale dell'opera applicando metodi di segmentazione del testo operistico (si veda [2]). Il titolo dell'opera musicale consente di identificare in modo univoco una composizione all'interno del catalogo facendo confluire la forma normalizzata in un *authority file*. A livello implementativo, la gestione di questa informazione può essere affidata al modello relazionale, oppure a quello dei linked data nel Web semantico. In una base di dati tradizionale un'entità che si presenti in più forme differenti, che per il trattamento computazionale mancano di correlazione, richiede una rappresentazione relazionale preordinata delle varianti, che può risultare gravosa. Più immediato è l'ecosistema del Web semantico, che rappresenta la realtà basandosi su ontologie formali e connette le informazioni attraverso Linked Open Data (LOD). Il modello relazionale richiede la preparazione di una struttura entro la quale inserire l'informazione, mentre il modello semantico viene strutturandosi con la stessa realizzazione del dataset e può per questo essere più conveniente. In un dataset di asserzioni basate su triple espresse nella forma di LOD secondo il Resource Description Framework, i vocabolari controllati ai quali le informazioni fanno riferimento [1], codificati secondo gli standard SKOS (Simple Knowledge Organization System), prevedono che si assegni semplicemente la proprietà `<skos:prefLabel>` alla forma preferita e la proprietà `<skos:altLabel>` a ciascuna delle altre (si veda [6]: 11, per una esperienza attuata nel dominio musicale). Un esempio pratico può essere dato partendo da un caso che si riscontra nella collezione di incisioni storiche "Mauro Benedetti" conservata presso il Dipartimento di Beni Culturali dell'Università di Bologna, sede di Ravenna. Prendendo dalla collezione di dischi d'opera tre esemplari di 78 giri (vd. Fig. 1), tutti con l'incisione del duetto di Violetta e Alfredo dal terzo atto della Traviata di Giuseppe Verdi, ma recanti sull'etichetta fisica uno il titolo "Parigi, o cara, noi lasceremo", un altro il titolo "Parigi o cara", e il terzo "Alfredo! Ah tu il vedesti" (l'incipit del recitativo), nel dataset rappresentato in forma semantica, si può scegliere di dare al primo titolo la proprietà `<skos:prefLabel>` mentre gli altri due avranno attribuita la proprietà `<skos:altLabel>`.

Figura 1

Una tale mappatura delle forme varianti con una forma preferita fornisce un'intestazione uniforme del contenuto, un punto di accesso che vale per ciascuna rappresentazione di uno stesso brano d'opera, indipendentemente l titolo riportato sull'etichetta. Automaticamente questo lavoro coincide con quel controllo di autorità che produce l'*authority file*, ma il Web semantico rende più immediata la relazione tra la forma preferita e le forme varianti. Ad esempio, non adotta strategie di ricerca basate sull'ordinamento alfabetico, il quale non avrebbe rapporti con il contenuto semantico della risorsa cercata.

3. DICHIARAZIONE DI APPARTENENZA

I dati relativi alle risorse così descritte possono essere ricollegati con le informazioni presenti nella base di conoscenza per il melodramma *Corago*[4], presentata in [3], costituita da un esteso dataset in Linked Open Data e da un modello semantico basato sulle ontologie CIDOC Conceptual Reference Model e Functional Requirements for Bibliographic Records object

[3] Cfr. Regole italiane di catalogazione: https://norme.iccu.sbn.it/index.php?title=Norme_comuni/Authority_file/Titoli_materiale_musicale/Registrazione_di_authority/Dati_specifici/Estratto e in dettaglio https://norme.iccu.sbn.it/index.php?title=Titolo_dell%27opera_musicale/Capitolo_4#4.5.3
[4] Corago è consultabile all'indirizzo: https://corago.unibo.it/

oriented (FRBRoo, confluita in Library Reference Model object oriented, LRMoo[5]), punti di riferimento per le *Digital Humanities*.
Una registrazione su disco, come quelle prese in esame nell'esempio della sezione precedente, si colloca nel modello concettuale FRBR a livello di *manifestazione*, mentre l'evento performativo è espressione dell'opera "La Traviata" che a sua volta è *opera* nel senso del primo dei quattro livelli (si veda [7]: 108) composta da vari testi: testo letterario, testo musicale, scenografie, costumi, indicazioni di regia e di prossemica.
Il collegamento a "La Traviata", identificata attraverso l'URI presente nella base informativa *Corago*[6], può avvenire nel Web semantico attraverso la proprietà partOf del linguaggio OWL (Ontology Web Language) `<owl:partOf>`. Sarebbe naturalmente possibile inserire direttamente nel proprio dataset l'opera "La Traviata" e collegare ad essa i dati relativi ai documenti correlati, per poi rinviare questa entità attraverso la proprietà sameAs, che si pone come una vera e propria dichiarazione di equivalenza (si veda [5]: 28), all'URI presente nella base informativa *Corago*. I rischi impliciti in un utilizzo poco consapevole di sameAs, spiegati dettagliatamente in [4], seppur ridotti quando ci si muove all'interno di domini specifici, portano a preferire l'uso di altri costrutti assertivi. Per restringere il campo e adottare ontologie specifiche del dominio dei beni culturali, si può naturalmente fare uso di LRMoo, che renderebbe possibile questo collegamento tramite la proprietà R67 hasPart[7]:
F1_Work (opera) R67_has_part F1_Work (aria)
Grazie a un collegamento di questo genere, si stabilisce una relazione che disambigua un concetto in modo semplice con economia in termini di tempo e di lavoro computazionale. Asserzioni più o meno elementari, strutturate in forma di triple, compongono il grafo di conoscenza che rappresenta la nostra descrizione di una parte dello scibile e successivamente vanno a collegarlo in modo immediato con entità facenti parte del grafo globale dell'informazione ognuna delle quali verificata e soggetta al controllo di autorità. La rete del Web semantico va arricchendosi progressivamente con l'inserimento di nuovi contenuti, arricchendo con nuova informazione i dati collegati.

4. GESTIONE DI INDICI ANALITICI

Un aspetto fondamentale di cui tenere conto per la modellazione della conoscenza relativa ai contenuti di registrazioni musicali, per loro natura opere che si realizzano su un asse temporale, è la segmentazione. "La Traviata" di Giuseppe Verdi, ad esempio, opera lirica in tre atti su libretto di Francesco Maria Piave, è suddivisa in 11 numeri musicali[8]. Il contenuto dell'opera può essere descritto identificando attraverso indicazioni numeriche le sezioni e riconducendo a tale scansione qualunque brano preso in considerazione.
A chi prenda in esame il repertorio operistico, oltre a un linguaggio tecnico per il quale la pratica e i dizionari di dominio aiutano a orientarsi, la critica musicale fornisce una metodologia di presentazione ordinata che consiste in una descrizione morfologica attraverso un sistema di numerazione annidata da attribuire alle diverse parti che compongono l'opera. Muoversi al suo interno può essere agevole se ci si affida a un indice analitico nel quale la numerazione individua le sezioni e permette di identificare i contenuti con precisione. Per tornare al duetto di Violetta e Alfredo in Traviata, atto terzo, secondo gli studi sulla partitura autoriale la parte in cui è incluso deve essere indicata con il numero 10. Tale numero individua una unità musicale e drammaturgica nella sequenza continua del contenuto (si veda [2]: 363-365). Si potrebbe scendere molto più in dettaglio, perché si tratta di un sistema di numerazione annidata che consente una granularità fine nella segmentazione analitica, ma qui non occorrerà farlo.
Una numerazione strutturata, riportata in un indice analitico ragionato, aiuta nel collegare le porzioni di un'opera e qualsiasi documento ad esse relativo con la scansione univocamente determinata alla quale esse sono pertinenti. Il brano riprodotto su una facciata di un disco 78 giri potrà essere ricollegato al numero che ne indica la posizione occupata nell'opera. Come gestire l'indice analitico nel Web semantico? Nella modellazione della base informativa, anche questa relazione potrebbe essere ipotizzata come una appartenenza, che può essere descritta in OWL attraverso asserzioni che usino nel predicato la proprietà `<owl:partOf>` oppure in LRMoo ricorrendo alla proprietà R67 hasPart.

5. CONCLUSIONI

La conoscenza del repertorio operistico storico presso il grande pubblico è ancora limitata anche a causa di una scarsa circolazione delle informazioni dovuta alla deteriorabilità e rarità dei supporti. Per valorizzare il patrimonio sonoro delle

[5] In proposito si veda: https://repository.ifla.org/bitstream/123456789/2217/1/144-riva-en-paper.pdf
[6] http://corago.unibo.it/opera/Z000028188
[7] Si rimanda nuovamente a: https://repository.ifla.org/bitstream/123456789/2217/1/144-riva-en-paper.pdf
[8] Per una definizione di numero musicale si faccia riferimento al "Piccolo glossario di drammaturgia musicale" di Lorenzo Bianconi e Giorgio Pagannone: http://box.dar.unibo.it/files/didattica/Glossario-di-Drammaturgia_musicale.pdf

registrazioni storiche di canto lirico e renderlo accessibile a un vasto pubblico può essere applicato il modello della conoscenza del Web semantico con il quale i contenuti sono proposti accedendo a tutte le potenzialità offerte dai Linked Open Data: collegati a ontologie esistenti di riferimento, possono essere utilizzati sia nella creazione di *authority file* che per il collegamento delle parti con l'opera. Inoltre, rappresentano una possibile metodologia per la descrizione morfologica consistente in un indice analitico ragionato che produce un indice ordinato dell'opera permettendo di collocare univocamente con precisione ogni documento ad essa riferito.
I vantaggi sono evidenti sia per la navigazione da parte di sistemi automatici, che ne sviluppa le opportunità di circolazione e diffusione, sia per la consultazione da parte degli utenti finali, studiosi o semplici curiosi.

6. RINGRAZIAMENTI

Si ringrazia cortesemente per le immagini il Laboratorio Musicale del Dipartimento di Beni Culturali dell'Università di Bologna che gestisce la ricca collezione discografica intitolata a Mauro Benedetti[9].

BIBLIOGRAFIA

[1] Berti, Michela, and Manuela Grillo. 'Strumenti digitali per lo studio delle arti performative storiche: il database e il thesaurus PerformArt'. *Umanistica Digitale* 9, no. 10 (9 September 2021): 443–450.

[2] Bianconi, Lorenzo, Angelo Pompilio, and Giorgio Pagannone. 'RADAMES. Prototipo d'un repertorio e archivio digitale per il melodramma'. *Il saggiatore musicale* XI (2004): 345–394.

[3] Bonora, Paolo, and Angelo Pompilio. 'Corago in LOD. The Debut of an Opera Repository into the Linked Data Arena'. *JLIS.It* 12, no. 2 (2021): 54–72.

[4] Halpin, Harry, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. 'When Owl:SameAs Isn't the Same: An Analysis of Identity in Linked Data'. *In The Semantic Web – ISWC 2010*, edited by Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, 305–320. Berlin, Heidelberg: Springer, 2010.

[5] Stahmer, Carl. 'Making MARC Agnostic'. In *Linked Data for Cultural Heritage*, edited by Ed Jones and Michele Seikel, 23–39. London: Facet Publishing, 2016.

[6] Thorsen, Hilary K., and M. Cristina Pattuelli. 'Linked Open Data and the Cultural Heritage Landscape'. In *Linked Data for Cultural Heritage*, edited by Ed Jones and Michele Seikel. London: Facet Publishing, 2016.

[7] Tomasi, Francesca. Organizzare la conoscenza: Digital Humanities e Web Semantico. Milano: Editrice Bibliografica, 2022.

[9] Per una presentazione della collezione si veda: https://site.unibo.it/collezioni-discografiche-dbc/it/collezione-benedetti

PUBLIC HISTORY E ARCHEOLOGIA DIGITALE

Archeologia e rilievo 3D: una riflessione sulle metodologie. Due casi studio di area mediterranea.

Graziana D'Agostino[1], Pietro Maria Militello[2]
[1] Dipartimento di Ingegneria Civile e Architettura, Università di Catania, Italia - graziana.dagostino@unict.it
[2] Dipartimento di Scienze umanistiche, Università di Catania, Italia - pietro.militello@unict.it

ABSTRACT
Il tema della documentazione nativa digitale, in ambito archeologico, riguarda una ampia categoria di prodotti: a microscala, il singolo manufatto; ad un livello più ampio, che potremmo definire di mesoscala, i modelli 3D di singoli monumenti; a livello ancora più ampio (macroscala), l'approccio territoriale. Il rilievo 3D di singoli monumenti ha ormai una consolidata tradizione. Esso costituisce un tipico caso di innovazione radicale apportato dagli strumenti computazionali, a differenza di altri strumenti che rappresentano sostanzialmente un potenziamento di un approccio tradizionale. Lo sviluppo di questo sistema ha introdotto una enorme varietà di attrezzature e software con una ampia gamma di costi e prestazioni, che pongono all'archeologo il problema della selezione. Allo scopo di costruire delle linee guida per l'archeologo si è voluto mettere a confronto metodologie diverse applicandole a tre casi studio. Il presente contributo, attraverso i casi studio indagati, si inserisce all'interno di una riflessione più ampia ed ancora attuale sul ruolo del digitale nel campo della documentazione e divulgazione archeologica.

PAROLE CHIAVE
Archaeology; digital survey; laser scanner; SfM; digital documentation.

1. INTRODUZIONE
Il tema della documentazione nativa digitale, in ambito archeologico, riguarda una ampia categoria di prodotti [14]: a microscala, il singolo manufatto; ad un livello più ampio, che potremmo definire di mesoscala, i modelli 3D di singoli monumenti [3]; a livello ancora più ampio (macroscala), l'approccio territoriale, come ad esempio le applicazioni di GIS 3D in ambito archeologico [12, 18]. Mentre queste ultime sono ancora in una fase di elaborazione, anche a seguito delle diverse tendenze nella gestione materiale dello scavo, il rilievo 3D di singoli monumenti, (da distinguere dalla ricostruzione 3D, nella quale il modello è ricostruito tridimensionalmente, seguendo un approccio scientifico, su programmi di modellazione 3D [13], ha ormai una consolidata tradizione. Esso costituisce un tipico caso di innovazione radicale apportato dagli strumenti computazionali, a differenza di altri strumenti che rappresentano sostanzialmente un potenziamento di un approccio tradizionale. Lo sviluppo di questo sistema ha introdotto una enorme varietà di attrezzature e software con una ampia gamma di costi e prestazioni, che pongono all'archeologo il problema della selezione degli strumenti. Non sono stati rari i casi in cui il prodotto commissionato a ditte specializzate soprattutto da enti, come le soprintendenze, è risultato sovradimensionato, se non inutilizzabile, in relazione agli obiettivi prefissi. La documentazione 3D può infatti assolvere a diverse funzioni: essere una forma di documentazione fine a sé stessa, da conservare in archivio; essere utilizzata per lo studio del monumento; fornire un supporto per le proposte di ricostruzione; essere utilizzata per la divulgazione o essere uno strumento di diagnostica dei cedimenti strutturali nel tempo. È inevitabile, anche se non sempre chiaro, che ognuna di queste esigenze presuppone approcci diversi, nei quali un peso non indifferente gioca il rapporto tra costi e risultati: il grado di precisione di una restituzione a scopi divulgativi può essere anche dell'ordine di centimetri, laddove una restituzione a scopo di monitoraggio necessita di precisione millimetrica. Soprattutto per chi lavora in istituzioni pubbliche con limitate disponibilità di denaro è importante avere idee chiare su quali metodologie richiedere e quali risultati aspettarsi in relazione agli scopi del lavoro. Allo scopo di costruire delle linee guida per l'archeologo si è voluto mettere a confronto metodologie diverse applicandole a due casi studio.
Il primo caso studio è il Quartiere Sud-Ovest del Palazzo di Festòs, a Creta (Grecia). Si tratta di un complesso architettonico, denominato Quartiere Levi dal nome dello scavatore, databile ad età protopalaziale (1950-1700 a.C.) formato da diversi ambienti, conservati in alcuni tratti fino al pavimento del secondo piano per una altezza massima di 6 metri [16]. A causa della deteriorabilità dei materiali impiegati, il Quartiere Levi non è accessibile ai visitatori ed è stato protetto da una copertura che ne rende impossibile il rilievo da drone. In questo caso, il nostro obiettivo era, in primo luogo, fornire una documentazione dello stato di fatto per confrontarlo, successivamente, con altre acquisizioni 3D e rilevare il livello di degrado di una struttura estremamente fragile. Ulteriore obiettivo era elaborare uno strumento per la fruizione di spazi

altrimenti non accessibili ai visitatori e, infine, integrare l'ortofoto del palazzo di Festòs realizzata in anni precedenti[1] nella quale del Quartiere è visibile solo la copertura.
Il secondo esempio qui indagato è quello del sito di Calaforno (Ragusa), una struttura ipogeica di 35 ambienti preceduti da un grande camerone, che si sviluppa per ca. 150 metri, databile all'Età del Rame tardo (2700-2200 a.C.) [15]. In questo caso, il rilievo 3D costituiva uno strumento fondamentale per la documentazione di una architettura astrutturale, priva di elementi significativi di discontinuità che sono quelli su cui si appunta il rilievo tradizionale. Meno importanti erano gli altri due obiettivi di documentazione dello stato di fatto e di costruzione di una visita virtuale in quanto il monumento non soffre di urgenti problemi di degrado.
In tutti e due i casi qui esposti, i differenti rilievi digitali sono stati integrati con rilievi topografici effettuati mediante la creazione di una rete di punti georeferenziati sul terreno (nel caso di Creta il sistema di riferimento è stato il Greek Geodetic Reference System GGRS 87 EGSA 87), e con rilievo fotogrammetrico da drone con calcolo della Average Ground Sample Distance (GSD) [4, 5, 11].

2. DOCUMENTAZIONE 3D IN ARCHEOLOGIA

Nel campo della documentazione digitale 3D, allo stato attuale, esistono differenti tecnologie e metodologie di acquisizione metrica tridimensionale (consolidate ed in fase di sperimentazione), che si differenziano tra loro per tipologia di strumentazione utilizzata, applicazione in campo e accuratezza e tipologia del prodotto ottenuto [7, 17, 20]. Le metodologie di rilievo digitale maggiormente impiegate risultano essere le tecniche TLS (Terrestrial Laser Scanner - che prevedono l'uso di Laser Scanner terresti fissi su treppiedi), le più recenti tecniche iMMS (indoor Mobile Mapping Systems - che prevedono l'uso di Laser Scanner che operano in movimento) e le tecniche SfM (Structure from Motion - che prevedono l'uso di una macchina fotografica o di un drone). Tutte le tecniche di rilievo, seppur con evidenti differenze tecniche riassunte nella tabella 1, consentono di ottenere un modello tridimensionale numerico dell'oggetto di studio (nuvola di punti), dotato di informazioni metriche e cromatiche, che può essere successivamente.

	Terrestrial LaserScanning	**Mobile Mapping System**	**Structure from Motion**
Strumentazione	laser scanner	laser Scanner	macchina fotografica
Tipologia	fisso su treppiedi	portatile	portatile o su treppiedi
Costo strumentazione	alto	alto	medio/basso
Prodotto ottenuto	nuvola di punti	nuvola di punti	nuvola di punti
Densità di punti	alta	bassa	alta
Dato cromatico	si (media risoluzione)	si (bassa risoluzione)	si (alta risoluzione)
Tempo di acquisizione	alto	basso/medio	medio/alto
Tempo di post-processamento	alto	basso	alto
Punti di forza	Accuratezza del dato, portata	Manovrabilità, flessibilità, velocità	Accuratezza del dato, manovrabilità, flessibilità

Tabella 1

La scelta del flusso di lavoro da adottare e delle strumentazioni più appropriate è strettamente condizionata, anzitutto, dalle peculiarità ambientali e morfologiche del sito da indagare e, soprattutto, da una scrupolosa valutazione preliminare riguardante la qualità della documentazione desiderata, gli obiettivi che intendono essere conseguiti nell'ambito di una ricerca o attività professionale, nonché dalla facilità con cui è possibile effettuare le misurazioni sul campo e elaborare i dati acquisiti. Da un lato, le sfide pratiche e l'esigenza di competenze adeguate motivano l'impiego di strumentazioni in grado di fornire un'ampia gamma di dati dimensionali e materici. D'altra parte, la sovrabbondanza di informazioni ottenuta richiede successivamente un'adeguata gestione, elaborazione e archiviazione, le quali non sempre sono sostenibili in termini di risorse da parte del committente e/o dell'operatore. L'impiego delle tecnologie digitali richiede competenze operative avanzate e, soprattutto, la capacità di ottimizzare i dati raccolti al fine di produrre risultati di alta qualità che siano utili per vari scopi scientifici in ambito culturale, come la documentazione, l'archiviazione, la rappresentazione, la valorizzazione e la fruizione virtuale.
In questo contesto, la ricerca ha sollevato considerazioni riguardanti le potenzialità offerte dalle più recenti apparecchiature e l'accessibilità, nonché l'utilità, dei dati tridimensionali acquisiti per la visualizzazione e la documentazione del patrimonio

[1] L'ortofoto è stata realizzata nell'ambito del Progetto Festòs, Università di Salerno, direttore Fausto Longo.

censito. Queste includono la possibilità di aggiornare la documentazione bidimensionale attraverso l'ottenimento di ortofoto e profili a qualsiasi quota e in qualsiasi momento senza la necessità di un ritorno sul sito, l'abilità di estrarre dati numerici che necessitano di informazioni tridimensionali, come il calcolo di superfici o volumi, e la possibilità di creare visite virtuali da remoto grazie alla creazione di un modello tridimensionale fotorealistico, potenzialità che permettono di rispondere a richieste diverse e stimolare nuove finalità di ricerca.
Sulla base di queste ampie riflessioni di cui il dibattito rimane sempre aperto ed in evoluzione, si è scelto di mettere a confronto gli iter metodologici attuati per l'acquisizione e la documentazione digitale di due siti archeologici, differenti per caratteristiche morfologiche ed obiettivi di documentazione prefissati, ma che hanno in comune condizioni ambientali ostative che hanno stimolato la sperimentazione e l'integrazione di differenti tecnologie applicate all'acquisizione 3D dei luoghi.

3. DUE CASI STUDIO DI AREA MEDITERRANEA

Il primo caso studio qui riportato come esempio applicativo della tematica discusa è un complesso architettonico all'interno dell'area archeologica di Festòs (Creta), non visitabile dai turisti a causa dello stato di conservazione delle strutture e dei materiali che lo costituiscono, il cosìdetto quartiere Levi. Per tali ragioni, al fine di salvaguardare l'integrità dell'area, è stata installata una copertura a protezione, la cui efficacia sia dal punto di vista funzionale che estetico risulta discutibile. Il sito è composto da una sequenza di spazi, talvolta a doppia altezza, caratterizzati da dimensioni relativamente contenute, interconnessi da corridoi stretti e aperture di varie grandezze. Una precedente campagna di rilevamento digitale, mediante l'impiego di tecnologia laser scanning, ha prodotto risultati insoddisfacenti, in quanto l'obiettivo era la creazione di un modello tridimensionale atto a fornire una vista nadirale, georiferita, della sommità delle strutture. A causa della limitata distanza a cui è stata posizionata la copertura rispetto alle strutture e delle dimensioni della strumentazione impiegata, il modello numerico ottenuto presentava numerose zone d'ombra. Si è deciso, quindi, di utilizzare la tecnica di rilievo Structure from Motion, vista la complessità delle condizioni ambientali in cui si trova il sito (vd. Fig. 1).

Figura 1. Quartieri Levi, Festòs (Creta). Strumentazioni utilizzate ed attività di rilievo in campo.

Sono stati condotti e sperimentati due diversi approcci di acquisizione fotogrammetrica: uno mediante l'utilizzo della action cam GoPro Black Hero 6 (dotata di un obiettivo che replica l'effetto fisheye, caratterizzato da un campo visivo più ampio rispetto alle lenti convenzionali), e l'altro mediante l'impiego di una fotocamera Canon EOS 70D equipaggiata con un obiettivo tradizionale. Il post-processamento dei dati, avvenuto all'interno del software Agisoft Metashape, ha prodotto, per entrambi i dataset, un modello poligonale fotorealistico di elevata qualità [6]. L'innovazione scientifica, in questo approccio, è stata l'uso della action cam (di facile manovrabilità per la sua leggerezza e dimensioni ridotte) che ha permesso di ottenere un modello tridimensionale completo e caratterizzato da dati cromatici di alta qualità, utile anche alla realizzazione di una visita virtuale da remoto per quella zona dell'area archeologica chiusa al pubblico. Inoltre, grazie al

modello 3D è stato possibile ottenere l'ortofoto zenitale dell'area, difficilmente ottenibile con altre strumentazioni o tecniche di rilievo 3D. A tal fine, la nuvola di punti fotogrammetrica è stata ripulita della tettoia sovrastante e successivamente convertita in mesh texturizzata, dalla quale è stato possibile ricavare il geotiff del quartiere, prodotto digitale richiesto per completare l'ortofoto dell'intera area archeologica (vd. Fig. 2).

Figura 2. Quartieri Levi, Festòs (Creta). A sinistra, vista prospettica della nuvola di punti finale ottenuta con l'action cam GoPro; a destra, Geotiff.

Il secondo caso studio riguarda un ambiente sotterraneo caratterizzato da un percorso irregolare di 100 m ca., costituito da una sequenza di 35 stanze e da un ingresso monumentale formato da un corridoio esterno (il cosiddetto dromos), che conduce a una sala d'ingresso principale (circa 12 x 4 m) scavata nella roccia. Tutte le camere sono a pianta circolare e hanno un diametro che varia da 1,5 a 3 m e altezza da 1,6 a 1,8 m, collegate tra loro da aperture molto strette e in condizioni di totale assenza di illuminazione. La particolare conformazione morfologica del sito, l'Ipogeo di Calaforno (RG, Italia) e le condizioni ambientali hanno suscitato l'interesse a testare l'efficacia delle seguenti tecniche e strumentazioni per il rilievo digitale: il laser scanner Leica Geosystem P30 per il rilievo dell'intero Ipogeo (campagna di rilievo 2017-2021), il laser scanner Leica BLK360 Imaging per le strutture esterne (campagna di rilievo 2021), l'action cam GoPro Black Hero 6 per il rilievo fotogrammetrico di uno dei vani di dimensioni più ridotte e il sistema di mappatura Mobile Laser Scanner basato su Simultaneous Localization and Mapping (SLAM) BLK2GO della Leica Geosystem per il rilievo dei 35 vani (vedi dettagli di acquisizione e post-processamento in tabella 2).

	P30	**BLK2GO**
Numero di punti	≈ 230 milioni	≈ 4,8 milioni
Densità di punti	High	Low
Scala di restituzione	1:20	1:100
Livello di rumore	Low	High
Tempo di acquisizione	12 ore	24 minuti
Tempo di post-processamento	8 ore	30 minuti
Peso del file (.e57)	6550 MG	94,4 MB

Tabella 2

Lo studio condotto sull'Ipogeo di Calaforno si è principalmente concentrato su tre distinte finalità orientate alla documentazione e alla valorizzazione del sito: 1) aggiornamento della documentazione grafica bidimensionale del monumento, 2) sperimentazione di una modellazione poligonale fotorealistica ad alta risoluzione per consentire la fruizione virtuale degli ambienti ipogeici, 3) creazione di un modello 3D, utile a differenti analisi e approfondimenti che il sito offre di testare all'interno della ricerca archeologica (vd. Fig. 3) [8, 9, 10, 19].

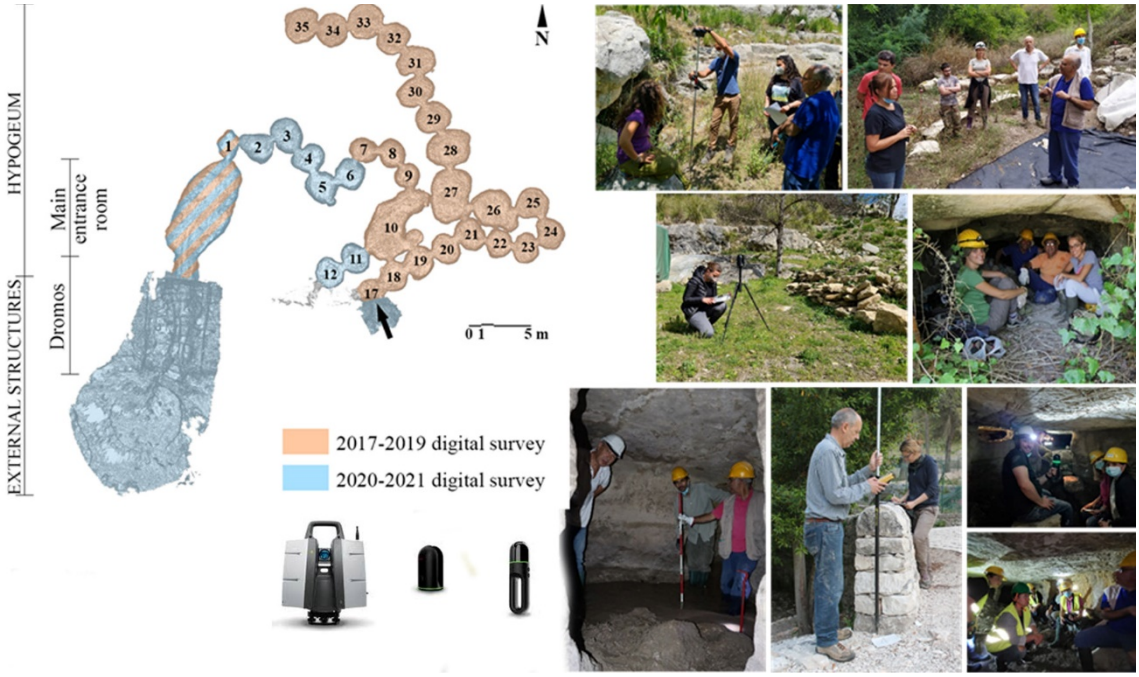


Figura 3. Ipogeo di Calaforno (RG, Italia). A destra, vista zenitale della nuvola di punti e strumentazioni utilizzate; a sinistra, attività di rilievo in campo.

La scelta di questo caso studio si è fondata su due criteri predominanti: l'importante valore storico-archeologico e culturale del sito e la presenza di fragilità intrinseche e criticità correlate a fattori morfologici e idrogeologici. Le condizioni ambientali complesse, con gli ambienti ipogeici totalmente al buio e i suoi 35 vani stretti e collegati, lungo un percorso a serpentina, da passaggi di dimensioni ridotte, hanno generato la necessità di documentare e comunicare i ricchi contenuti culturali intrinseci del sito, data anche la sua difficile accessibilità. In questa particolare area, l'utilizzo del rilievo digitale ha dimostrato di essere di fondamentale importanza e di grande efficacia nel restituire la complessa morfologia del sito. Questo perché, oltre a essere caratterizzato da ambienti stretti e bui, presenta anche un significativo dislivello tra i due punti di accesso e la parte finale del sito. Pertanto, la possibilità di sezionare il modello digitale a diverse quote ha consentito di restituire una documentazione bidimensionale estremamente accurata del sito (vd. Fig. 4).

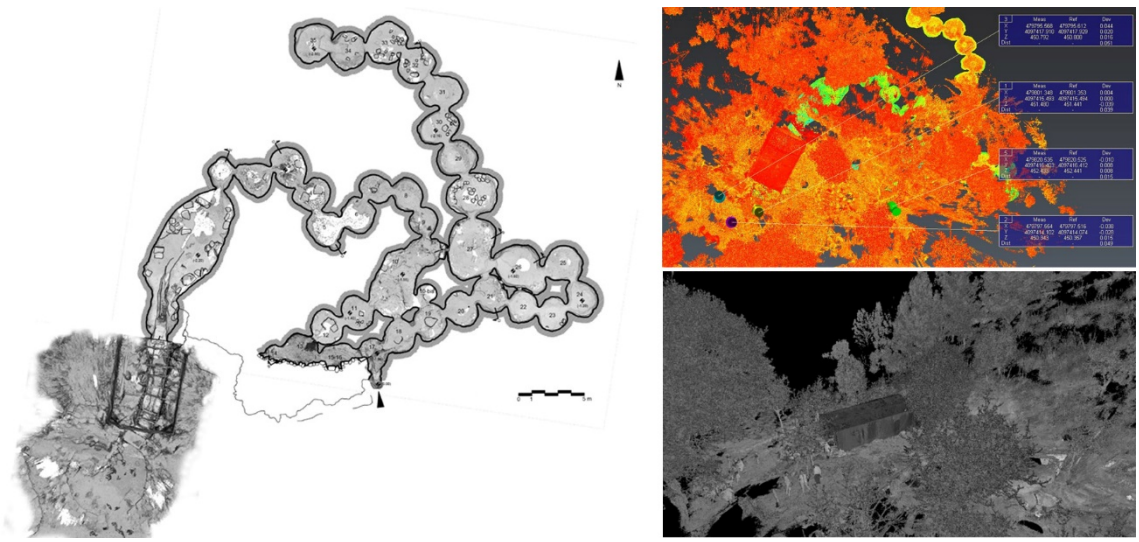


Figura 4. Ipogeo di Calaforno (RG, Italia). A sinistra, Nuova rappresentazione 2D planimentrica dell'Ipogeo di Calaforno, con ortofoto ricavate dalla nuvola di punti finale; a destra, vista prospettica della nuvola di punti dell'ambiente esterno.

La ricerca ha altresì fornito l'opportunità di mettere a confronto le diverse apparecchiature di rilievo, valutandone la qualità e l'accuratezza dei dati, nonché i benefici intrinseci derivanti dall'impiego di una determinata tecnologia rispetto ad un'altra, considerando le peculiarità del sito e gli obiettivi di documentazione [1, 2]. Il modello 3D dell'Ipogeo ha stimolato nuove

prospettive di ricerca, oltre alla possibile fruizione virtuale e all'aggiornamento della documentazione grafica. Le possibilità di studio si ampliano notevolmente: è ora possibile esaminare dettagliatamente le superfici dell'ipogeo, anche da un semplice computer, per individuare anomalie o tracce di lavorazione. Quest'analisi risulta essere molto più agevole rispetto all'esame in loco, spesso reso difficile dall'assenza di luce. Inoltre, è stato possibile estrarre parametri metrici come la cubatura degli spazi ipogeici, un'indagine difficile da eseguire con i metodi tradizionali. Ciò ha permesso osservazioni sulla quantità di materiale rimosso e sul lavoro impiegato, consentendo di verificare ipotesi sulla gerarchia e la cronologia dei vani.

4. CONCLUSIONI

Con questo contributo abbiamo inteso mettere a disposizione l'esperienza di due casi studio di area mediterranea, differenti per contesto e conformazione morfologica, ma similari per condizioni ambientali ostative, con conseguenti approcci metodologici e pratici differenti, utili a contribuire al dibattito sempre in atto tra il miglioramento tecnologico delle metodologie e le concrete esigenze Da questo punto di vista, i metodi di indagine e documentazione bidimensionale e tridimensionale proposti e sperimentati nei casi studio precedentemente delineati si fondano su un continuo dialogo tra ingegneri ed archeologi. Essi dimostrano che non esiste una soluzione migliore delle altre, ma la scelta va valutata in funzione dei contesti, degli obiettivi, basandosi su diversi parametri. Il primo è la tipologia di manufatto/monumento, per cui, ad esempio, una struttura architettonica templare richiede una documentazione accurata al millimetro di scala di cui un'architettura astrutturale non necessita. Il secondo parametro è l'uso ultimo che ne deriva: divulgazione virtuale e/o documentazione bidimensionale correlata da elaborati ottenuti dal modello digitale, quali base d'appoggio per interventi di restauro o diagnostica del degrado nel tempo. Un terzo parametro è l'accessibilità del monumento e le condizioni di rischio eventualmente presenti che suggeriscono l'uso di strumenti di esecuzione rapida, anche a costo di perdita di qualità (come, ad esempio, uno scavo in grotta per cui risulta utile l'uso della tecnologia SLAM). Un quarto parametro è sicuramente il costo economico e lavorativo rappresentato anche dalle competenze e dal tempo di post-processamento. L'approccio interdisciplinare alla ricerca ha portato alla riflessione sulle opportunità offerte dalle moderne apparecchiature e sulla comprensione e l'utilità dei dati tridimensionali per la visualizzazione e la documentazione del patrimonio storico. Riflessione possibile mettendo in continua relazione, da una parte, il rilievo tradizionale che si fonda sull'interpretazione esperta del rilevatore o dello studioso (rappresentazioni grafiche di alta qualità utilizzate per la documentazione scientifica), e dall'altra parte, il rilievo digitale 3D, il quale, anch'esso guidato dalle capacità interpretative del rilevatore, risponde a diverse esigenze aprendo a nuove prospettive la ricerca.

5. CREDITI E RINGRAZIAMENTI

Pietro Maria Militello ha scritto il paragrafo 1. Graziana D'Agostino ha scritto i paragrafi 2 e 3. Entrambi gli autori hanno scritto il paragrafo 4.

Il confronto tra le differenti attività di ricerca è stato condotto all'interno del progetto interdipartimentale "Storage. Dai dati al Web" (programma Pia.Ce.Ri.) dell'Università di Catania (2020-2024).

Le attività di rilievo in campo svolte con il laser scanner Leica P30 sono state eseguite con il contributo tecnico di Antonio e Salvatore Garro (3D Dimension Company, Catania, Italia). Le attività di rilievo in campo eseguite con il laser scanner BLK2GO sono state eseguite con il contributo tecnico di Nicolò di Blasi (Agente di Commercio Leica Geosystem) e Rosario Caruso (consulente tecnico di Leica Geosystem). Il laser scanner BLK360 è una strumentazione del Laboratorio di Fotogrammetria Architettonica e Rilievo "Luigi Andreozzi", sezione del Laboratorio RDA - DICAR - Università di Catania (Responsabile: Prof.ssa Mariateresa Galizia).

Si ringraziano per i permessi accordati: la Scuola Archeologica Italiana di Atene e l'Eforia alle antichità di Herakleion, Creta (Festòs) e la Soprintendenza BBCCAA di Ragusa (Calaforno).

BIBLIOGRAFIA

[1] Aiello, Damiano, Alessandro Basso, Maria Teresa Spena, Graziana D'Agostino, Umberto Montedoro, Mariateresa Galizia, Rosario Grasso, e Cettina Santagati. «The Virtual Batcave: a project for the safeguard of a UNESCO WHL fragile ecosystem». *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLII-2/W9 (2019): 17–24.

[2] Arena, Marinella, Daniele Colistra, e Domenico Mediati. «La grotta degli asceti. Rilievo e analisi dell'eremo di Santa Maria della Stella/ The Cave of the Ascetics. Survey and Analysis of the Hermitage of Santa Maria della Stella». a cura di Mirco Cannella, Alessia Garozzo, e Sara Morena, 753–76. Franco Angeli, 2023.

[3] Boyd, Michael, Rosie Campbell, e Roger Doonan. «Open Area. Open Data: Advances in reflexive Archaeological Practiece». *Journal of Field Archaeology* 46, fasc. 2 (2021): 62–80.

[4] Buscemi, Francesca. «Festòs 2014. L'attività di rilievo digitale». In *L'attività dell'università di Catania a Festòs nel 2013-2014. Annuario della Scuola Archeologica Italiana di Atene*, a cura di Francesca Buscemi e Pietro Maria Militello, 293–302. ASAA XCIII, III, 15, 2015.

[5] Buscemi, Francesca, e Marianna Figuera. «Managing complexity in long-term excavations: the GIS of Phaistos». In *GIS In Crete: archaeological questions and computational answers. Athens May 30-31 2024*, a cura di Vyron Antoniadis e Quentin Drillat, in corso di stampa.

[6] Buscemi, Francesca, Pietro Maria Militello, Cettina Santagati, Damiano Aiello, Graziana D'Agostino, e Marianna Figuera. «Use and reuse of spatial and quantitative data in archaeology: from 3D survey to serious game at Phaistos (Crete)». *Archeologia e Calcolatori* 31 (2020): 189–212.

[7] Calvano, Michele, Luciano Cessari, e Elena Gigliarelli. «Tradition in Innovation. Some Considerations on SLAM Technique Integration for Historic Buildings». In *Transizioni. Atti del 44° Convegno Internazionale dei Docenti delle Discipline della Rappresentazione/Transitions. Proceedings of the 44th International Conference of Representation Disciplines Teachers. Palermo 14-15-16 settembre 2023*, a cura di Mirco Cannella, Alessia Garozzo, e Sara Morena, 2521–30. Franco Angeli, 2023.

[8] D'Agostino, Graziana, Marianna Figuera, Valeria Pennisi, Gloria Russo, Mariano Sanfilippo, Pietro Maria Militello, e Rosaria Ester Musumeci. «Hydraulic risk Assessment in archaeological sites supported by an integrated digital survey – CFD (Computational Fluid Dynamics) monitoring approach». *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLVI-M–1 (2021).

[9] D'Agostino, Graziana, Marianna Figuera, e Gianluca Rodonò. «Digital survey and reception structures for a virtual fruition: the case study of the Hypogeum of Calaforno (Ragusa)». In *Proceedings of the joint international event 9th ARQUEOLÓGICA 2.0 & 3rd GEORES, Valencia (Spain) 26-27-28 aprile 2021*, 569–72. València: Editorial Universitat Politècnica de València, 2021.

[10] D'Agostino, Graziana, Marianna Figuera, Gloria Russo, Mariateresa Galizia, e Pietro Maria Militello. «Integrated 3D survey for the documentation and visualization of a rock-cut Underground Built Heritage». *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLVI-2/W1 (2022): 167–74.

[11] De Guidi, Giorgio. «Quality Report Calaforno Full». In *Calaforno I*, a cura di Pietro Maria Militello. Oxford, 2024.

[12] Dell'Unto, Nicolò, e Giacomo Landeschi. *Archaeological 3D GIS*. London: Routledge, 2022.

[13] Demetrescu, Emanuel, e Daniele Ferdani. «From Field Archaeology to Virtual Reconstruction: A Five Steps Method Using the Extended Matrix». *Appl. Sci.* 11 (2021): 1–23.

[14] Gordon, Jodi M., Erin W. Averett, e Derek B. Counts. «Mobile Computing in Archaeology: Exploring and Interpreting Current Practices». In *Mobilizing the Past for a Digital Future*, 1–30. Grand Forks: Digital Press, 2016.

[15] Guzzardi, Lorenzo. «Un ipogeo preistorico a Calaforno e il suo contesto topografico». *Sicilia Archeologica* 42 (1980): 67–94.

[16] Levi, Doro. *Festòs e la Civiltà Minoica*. Vol. I. Roma: Edizioni dell'Ateneo, 1978.

[17] Mandelli, Massimiliano, Francesco Fassi, Luca Perfetti, e Carlo Polari. «Testing different survey techniques to model architectonic narrow spaces». *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLII-2/W5 (2017): 505–11.

[18] Nicoletti, Rossella. «Un progetto GIS per la mappatura, la gestione e la valorizzazione del patrimonio archeologico di Assoro (Enna)». In *Diari di un Archeologo*, a cura di Girolamo Sofia, I:100–118. Terme Vigliatore: Giambra Editori, 2022.

[19] Pennisi, Valeria, Graziana D'Agostino, Marianna Figuera, Gloria Russo, Mariateresa Galizia, Pietro Maria Militello, e Rosaria Ester Musumeci. «Hydraulic preservation of archaeological hypogeum based on digital survey techniques». In *Proceedings MetroArchaeo, International Conference on Metrology for Archaeology and Cultural Heritage. Cosenza 19-20-21 settembre 2022*, 96–100. IMEKO, 2022.

[20] Pietroni, Eva, e Daniele Ferdani. «Virtual Restoration and Virtual Reconstruction in Cultural Heritage: Terminology, Methodologies, Visual Representation Techniques and Cognitive Models». *Information* 12, fasc. 167 (2021): 1–30.

Dal Catasto Borbonico alla Genomica. Piattaforme digitali e interdisciplinarità: il progetto “We Are What They Were” di Riposto

Salvatore Spina
Dipartimento di Scienze Umanistiche, Università di Catania, Italia - salvatore.spina@unict.it

ABSTRACT

Il Mediterraneo è una dimensione-concetto – definito da Braudel «continente liquido» –, che origina e si definisce nel ruolo delle genti, delle culture e delle economie che si sono susseguite nei *suoi* tempi. Oggi, attraverso un approccio interdisciplinare, anche la ricerca genetica riesce a ridefinire questo ‘continente’, grazie al venire in essere di quel ponte dialettico tra la Storia e le Scienze Biologiche. Su questo assunto, trova ragione il progetto “We Are What They Were”, che guarda alla ricostruzione storica della comunità ripostese, tra Sette e Ottocento, attraverso un lavoro di interconnessione tra fonti primarie e dati genomici, i cui risultati confluiscono nella progettazione di un portale web, visto come strumento storiografico. In questo studio, l’approccio genealogico e l’analisi genetica emergono quali strumenti strategici per la descrizione della comunità, consentendo un workflow storico-metodologico focalizzato sull’analisi del cromosoma Y di un vivente, che, completando la storia dell’ascendenza patrilineare del cognome «Sorbello», partendo dal borgo ripostese, ha consentito di gettare le basi per una metodologia in grado di interconnettere dati e informazioni storiche, allo scopo di spiegare gli assetti delle comunità, la cui esatta configurazione è determinante per la descrizione dell’Europa e del Mediterraneo, in età moderna.

PAROLE CHIAVE

Big Data; Visualization; patrilineare; DNA antico.

1. INTERCONNESSIONI

Se vero che il Mediterraneo è uno “spazio” storico, è ancor più vero che tale concetto si definisce attraverso le genti, le culture, le economie e i meccanismi di iter-socializzazione che lo hanno reso un «continente liquido» [10,11,12]. E in un mondo sempre più interconnesso, tale “spazio liquido” si può spiegare attraverso una reale interdisciplinarità, che porta dentro la narrazione della storia anche gli «atomi e [i] geni» [39]. Se vero, poi, che tale intuizione ha avuto una prima teorizzazione già nei lavori di Cavalli-Sforza [15, 16, 17, 18, 19] e di Sorre [40], è con l’avanzamento nella ricerca genetica e genomica che le prove date dalle Scienze Naturali hanno potuto ridefinire la Storiografia e le sue metodologie. Così, il ‘continente liquido’ assume ulteriori significati, che non si limitano allo spazio e alle interrelazioni costruite sugli “affari”, ma è *luogo* che interagisce con l’uomo e con le sue strutture biologiche.

Il Mediterraneo preesiste all’uomo, ma diventa “problema storico” quando questi lo nomina, imponendogli un significato – su cui, oggi, tra l’altro, comincia a farsi sempre più viva una *storia ambientale* [2, 3, 4, 34, 35].

La Storia ha ‘significato’ i luoghi, i quali forniscono il primo *utile* e ricevono quel *disavanzo* che necessariamente li modifica. E ogni scambio tra uomo e ambiente, tra uomo-uomini e Mediterraneo, si è tradotto in un complesso documentario che ha testimoniato la Storia d’Europa. Ma se gli archivi e le scritture rappresentano il *passato*, il *Futuro* sta nel più grande database che la Storia abbia mai creato: il patrimonio genetico. In esso confluiscono tutti i dati e le informazioni in grado di spiegare la Storia dell’uomo. Gli archivi, infatti, in relazione a tale prospettiva, rappresentano l’espletamento dell’azione dell’uomo, mentre il patrimonio genetico è il paradigma che spinge all’azione, sta *a priori* d’essa; e la possibilità di creare un legame diretto tra le due dimensioni – l’archivio *a priori* (banche dati genomiche, come LocusLink, RefSeq, COGs, GeneCards) e quello *a posteriori* (gli archivi storici) –, potrebbe garantire la risposta alle storie e alla Storia, in chiave scalare e globale.

Sicché, se, nel corso dei secoli, non abbiamo mai avuto alcun ripensamento sul fatto che la documentazione scritta sia la prova unica dell’azione dell’uomo, oggi, questi ulteriori e potenti “Big Data della Storia” hanno la necessità di interconnettersi anche con le banche genomiche, al fine di descrivere aspetti delle comunità mediterranee, le quali hanno consentito a quel “continente” di assumere quella “mobilità” che lo ha reso spazio unico e identitario.

Nasce così il progetto/laboratorio “We Are What They Were”, che mira alla ricostruzione dell’assetto della comunità ripostese nel Settecento, attraverso una metodologia che vuole far dialogare le fonti primarie (informazioni catastali,

anagrafiche, notarili, amministrativi) con quelle derivanti dalle banche dati genomiche, allo scopo di creare un portale web che possa "linkare" le informazioni storiche a quelle genetiche e alla documentazione che "narra" le città, ossia i catasti.

2. "WE ARE WHAT THEY WERE"

I catasti rappresentano lo strumento migliore per il controllo della razionalità dell'urbanizzazione. Inizialmente concepiti per finalità urbanistiche e fiscali [23, 48], hanno assunto, nel tempo, una connotazione storica fortemente scientifica, facendosi fonte di una metodologia che è in grado di spiegare, attraverso di essi, cosa è l'uomo nello spazio; le sue relazioni economiche e le sue azioni volte alla creazione di un quadro urbano in grado di rispondere alle necessità di una comunità. Grazie ad essi, oltre alla definizione del disegno della città, è possibile muoversi verso un approccio che può "democraticizzare" la narrazione della Storia di una società, dentro la quale si trovano tutti i soggetti coinvolti nella sua costruzione. Questo processo, negli ultimi decenni, ha portato ad un ulteriore avanzamento nell'acquisizione dei dettagli, soprattutto quando essi si fondono con la Genealogia degli individui che hanno lasciato la loro "testimonianza" in quelle che vengono definite *fonti primarie*.

Collocare le persone negli spazi e attribuire loro una identità, attraverso le genealogie e la Genomica, rappresenta una prospettiva di ricerca storica che ha dimostrato come possono ampliarsi le visioni della ricostruzione di svariati aspetti che caratterizzano una società del passato, grande o piccola, come nel caso, ad esempio, del progetto di ricerca che ha visto il coinvolgimento di diversi ricercatori nello studio dei reperti ossei ritrovati a Roccapelago (Reggio Emilia) [21].

La Storia, quindi, può avvalersi di nuove fonti, di nuovi beni culturali [33] e di nuove informazioni. Dal canto suo, successivamente, la rivoluzione digitale, e la possibilità di sfruttare nuovi sistemi per la creazione di archivi digitali, offre la possibilità di far comunicare queste fonti, garantendo l'opportunità di mettere su mappe le informazioni storiche, sfruttando soprattutto i mappali a corredo dei Catasti, consentendo di operare una narrazione "visual" dell'assetto di una comunità, nello spazio e nelle relazioni con gli altri membri.

Sulla spinta di queste intuizioni, nel 2019 è stato avviato un progetto di ricerca, che ha fatto di Riposto il laboratorio [44] di un approccio multidisciplinare. L'obiettivo è quello di ricostruire l'assetto di questa cittadina, nel Settecento; secolo in cui, questo luogo, la cui storia si radica nella Contea di Mascali, ha visto una compenetrazione da parte di *coloni* [1, 24] provenienti da diverse zone della Sicilia, attirati dalla prospettiva di un potenziamento della struttura commerciale legata alla rada, da sempre porto naturale della Contea (e dell'Etna), vista come opportunità – soprattutto da parte del baronaggio acese [42, 43] – per la creazione di una piazza di scambio alternativa a quella catanese, la quale aveva subìto la sventura del terremoto del Val di Noto [28, 37].

Da un altro lato, tale progetto vuole fornire tutte le informazioni necessarie per la descrizione di una comunità marinara e mediterranea [41, 42, 43], allo scopo di spiegare e individuare le interconnessioni sociali, economiche, famigliari, culturali, gli standard e lo stile di vita tipiche di una società interconnessa con i luoghi del Mediterraneo.

Leitmotiv diviene, necessariamente, la costruzione di una piattaforma in grado di legare questi dati storici, non già nella prospettiva di un *invented archive* [22, 36, 46], quanto nella visione della realizzazione di un portale web che consenta una "narrazione" di tali documenti in chiave "visual" [5, 8, 25, 45], in grado di catalizzare in un unico "strumento storiografico" la descrizione di come una comunità si sia stanziata in uno spazio – inteso come *luogo* di uomini, famiglie ed affari – 'locale' e 'glocale'.

Il lavoro, nel 2021, diventa una Local Time Machine, in seno al più ampio progetto europeo "Time Machine Europe"[1], con il titolo "Il Catasto Borbonico in Sicilia (1845). Studio della sezione "Riposto"[2], inserendosi, così, all'interno della rete dei progetti digitali europei, quale ulteriore tassello del TME, che ambisce alla realizzazione di un network di prodotti di ricerca storica fondati su un approccio interdisciplinare.

Per "collocare" i soggetti registrati nel *Sommarione* della sezione «Riposto» del Cessato Catasto Borbonico, è stato utilizzato il tool "StoryMapJS"[3] (vd. Fig. 1), sviluppato dal Knight Lab della Northwestern University, grazie al quale è possibile *narrare* gli spazi [5, 8, 25, 45] prettamente storica.

Presso il Centro Regionale per l'inventario, la Catalogazione e le Documentazione grafica, fotografica, aerofotogrammetrica, audiovisiva, alla sezione *Catasto Borbonico Archivio Mortillaro di Villarena (1837-1853)*, è stato individuato il mappale «Giarre-Riposto»[4], nel quale sono stati localizzati i vari soggetti registrati nel *Sommarione* della sezione catastale di Riposto –conservato, questo, presso l'Archivio di Stato di Catania, dove, tra l'altro, sono stati

[1] https://www.timemachine.eu/time-machine-organisation/
[2] https://www.timemachine.eu/ltm-projects/il-catasto-borbonico-in-sicilia-1845-studio-della-sezione-riposto/
[3] https://storymap.knightlab.com
[4] Centro Regionale per l'inventario, la Catalogazione e le Documentazione grafica, fotografica, aerofotogrammetrica, audiovisiva, *Catasto Borbonico Archivio Mortillaro di Villarena (1837-1853)*, *Catania, Mappa del territorio di Giarre e Riposto (/Milo/Sant'Alfio/Santa Venerina/Zafferana Etnea)*, 121 (link: https://www.cricd.it/pages.php?idpagina=303).

individuati altri disegni che hanno permesso di ampliare il dettaglio delle informazioni relative alle varie zone del territorio del borgo marinaro (vd. Figg. 2, 3).

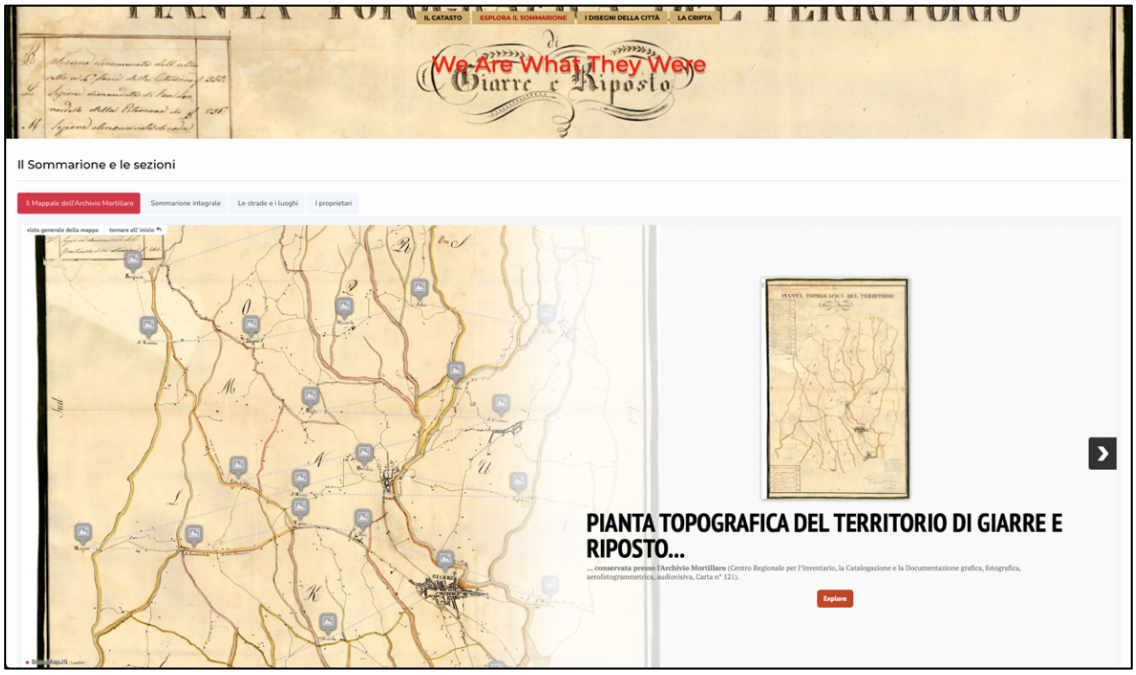


Figura 1

Figura 2

Su un altro piano, il lavoro prettamente storico-genealogico ha portato alla costruzione degli alberi delle famiglie più importanti di Riposto. Al momento, come già detto, è possibile consultare quelli della famiglia Calì, Fiamingo e Sorbello. Partendo dai cognomi presenti nel Catasto, si è cercato di arrivare al *match* tra questi e quelli dei *Registri di Stato Civile*, dei *Riveli* [26, 32], dei fondi notarili conservati presso l'Archivio parrocchiale della Chiesa San Pietro di Riposto, presso l'Archivio di Stato di Catania e quello di Palermo.

Ad oggi, il progetto si è focalizzato sulla collocazione delle genealogie di alcune delle famiglie più importanti della comunità ripostese (vd. Fig. 4), costruite su base patrilineare (o cognomiali); scelta metodologica che, da un lato, ha garantito la correlazione esatta tra i dati delle fonti primarie ed il soggetto, e, da un altro lato, ha consentito di preservare l'identità personale e famigliare dell'individuo.

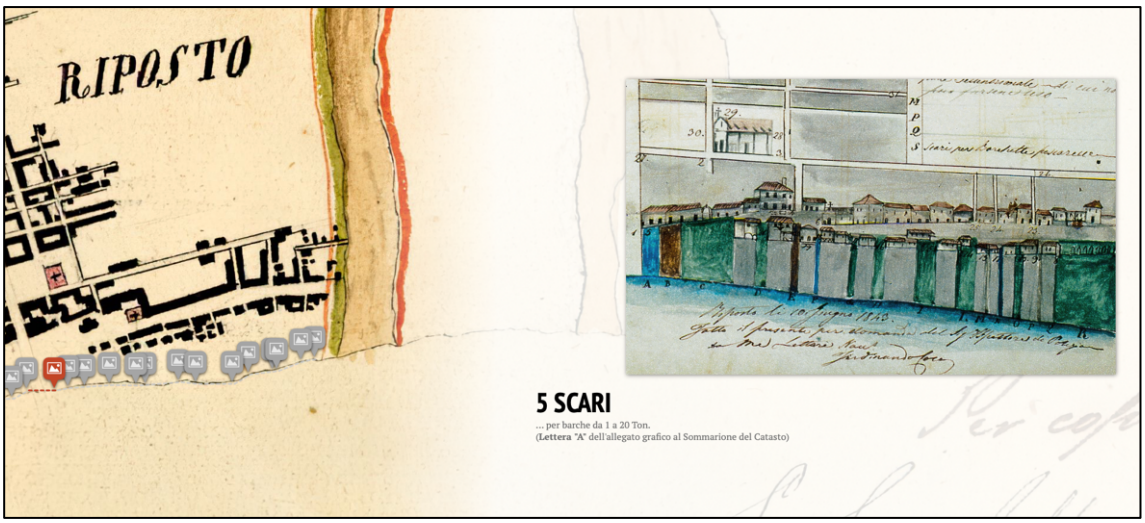


Figura 3

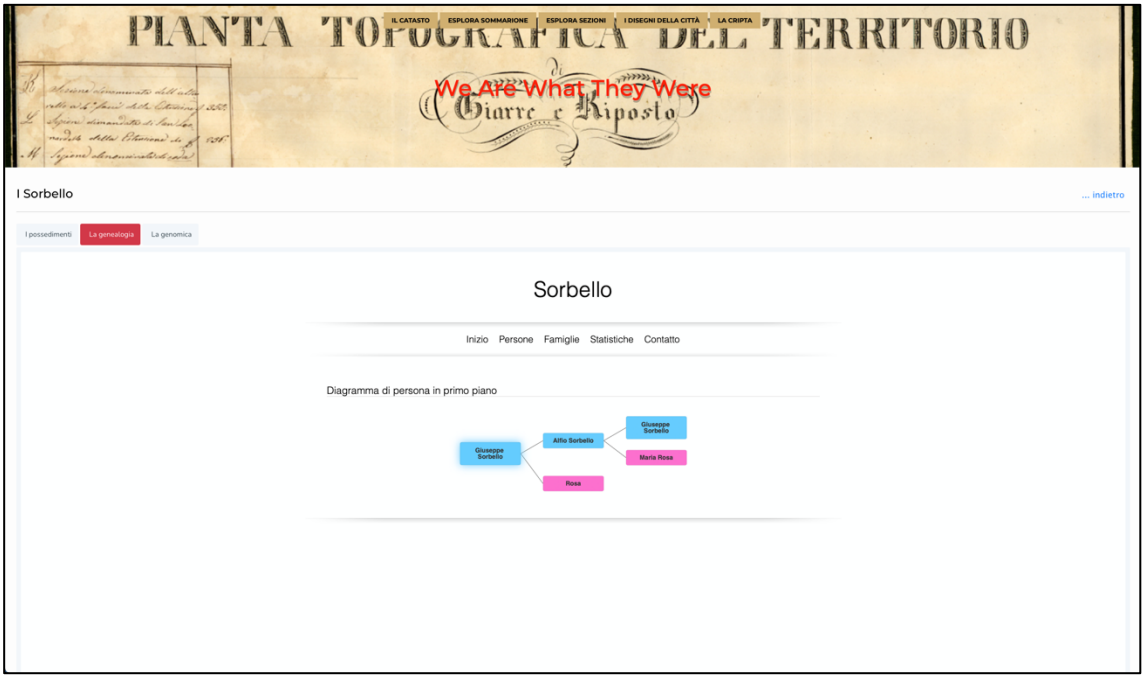


Figura 4

Le linee genealogiche cognomiali, infatti, emergono come una modalità efficace di rappresentazione e preservazione dell'identità attraverso il tempo, contrariamente alla raffigurazione a ventaglio dell'albero genealogico tradizionale, che non consente di garantire la chiarezza dell'identità individuale, soprattutto quando ci si spinge sempre più indietro nel tempo.

La sfida evidenziata è quella di tracciare la linea temporale di un gruppo di famiglie con un'ascendenza comune, mantenendo intatta l'identità di ciascuna di esse. La riflessione critica, su come delineare questa intricata rete genealogica, conduce all'idea che l'organizzazione in cognomi si configuri come l'approccio più idoneo per preservare e navigare la storia condivisa di tali famiglie. L'approccio cognomiale diventa, quindi, un veicolo essenziale per tracciare le connessioni delle famiglie nel corso dei secoli.

La ricerca sui documenti d'archivio (relativi alla persona), tuttavia, sebbene possa fornire un quadro storico dettagliato risalente anche fino al XIV secolo, inevitabilmente incontra limitazioni dovute all'assenza di documenti ancora precedenti. Tale vuoto, secondo le intuizioni dei già citati Cavalli-Sforza e Sorre – ma anche di una comunità scientifica che ha mostrato le possibilità dell'analisi genetica [7, 30, 31, 47] – può efficacemente essere colmato dall'analisi genetica. Questa metodologia supera le barriere delle lacune documentali, aprendo nuove prospettive di comprensione sulla formazione e l'evoluzione delle identità, di cui Riposto vuole essere esempio d'indagine.

L'analisi genetica, attraverso lo studio del cromosoma Y di un individuo, si presenta come una potentissima chiave per svelare la storia della sua ascendenza patrilineare. Il cromosoma Y, essendo ereditato di padre in figlio maschio, costituisce un percorso genetico che segue la linea maschile di discendenza. Questo aspetto si correla strettamente con l'eredità del cognome, il quale, di norma, si tramanda lungo la stessa linea patrilineare.
Attraverso la decodificazione degli SNPs (Single Nucleotide Polymorphisms [6, 29, 47]) mediante l'analisi del cromosoma Y, è possibile tracciare una narrativa genetica che si estende ben oltre la portata temporale dei documenti storici. Gli SNPs permettono di ricostruire le migrazioni e le connessioni ancestrali in modi che la documentazione tradizionale potrebbe non rendere accessibili.
Dal 2023, il portale[5] – in costante fase di sviluppo – diventa uno strumento storiografico che vuole spiegare le migrazioni interne al Mediterraneo e alla Sicilia, facendo di tale approccio un ulteriore strumento per definire le scelte economiche e d'interesse, tipiche di quella borghesia che costruisce il suo network, soprattutto attraverso le politiche matrimoniali, le quali furono determinanti non solo per la classe aristocratica [38].

3. LA GENOMICA. IL PRIMO APPROCCIO

Quando la ricerca archivistica incontra "legittime" limitazioni, le indagini biologiche possono colmare il vuoto tra i documenti ufficiali e l'individualità genetica, costruendo un ponte tra "resti" e riscontro notarile. Questo principio, oramai paradigmatico nella ricerca antropologica e archeologica, ha fatto sì che si guardasse alla possibilità di integrare le narrazioni *visual* delle famiglie della comunità con le informazioni derivanti dalle indagini genetiche effettuate sui viventi, facendo proprio il concetto di considerare la comunità dei vivi quale risultato genetico di quella dei secoli precedenti –We Are What They Were!
Così, il sistema collettivo di informazioni digitali del portale ha previsto la creazione di una scheda specifica in cui raccogliere i dati derivanti dall'analisi del DNA dei soggetti viventi.
Al momento, l'unico caso inserito nel portale riguarda la genomica relativa al cognome "Sorbello" (vd. Fig. 5), la quale, se da un lato ha consentito di "mappare" l'aplogruppo del soggetto (e dei suoi ascendenti) in un contesto euro-mediterraneo [44], da un altro lato, ha aperto il progetto verso la prospettiva di un coinvolgimento dell'intera comunità, allo scopo di intercettare i vari dati biologici, per definire le dinamiche della migrazione che ha portato alla conformazione della comunità ripostese. All'interno della scheda "Genomica", infatti, è possibile visualizzare la posizione del soggetto all'interno di un albero filogenetico, di identificare il suo aplogruppo Y e le subcladi. La pagina web relativa a queste ultime, ad es. R-Y133731 (del cognome "Sorbello"), descrive la sequenza di aplogruppi e le mutazioni che hanno portato alla formazione della subclade in oggetto.

Figura 5

[5] https://catastoriposto1845.altervista.org

È possibile, inoltre, ottenere una classe di informazioni ancora più dettagliata, cliccando sul link "*info*" dell'antenato comune (MRCA). Al suo interno è possibile trovare: 1) il nome dell'aplogruppo e maggiore dettaglio sull'età stimata di quest'ultimo (in alto a destra); 2) la formula utilizzata per il calcolo dell'età stimata dell'aplogruppo, che consiste in una media aritmetica arrotondata dell'età di ogni ramo; 3) la scheda TMRCA consente di consultare la lista dei rami che appartengono all'aplogruppo o alla subclade in oggetto; 4) la scheda SAMPLE, che include i dettagli relativi ai campioni di DNA sequenziato; 5) la scheda MAP, dove si evidenzia graficamente la localizzazione geografica dei campioni. Scorrendo la scheda verso il basso è possibile visionare la migrazione della linea dall'«Out of Africa» ad oggi, cliccando sul link "*Theoretical Computed Paths*".

4. L'OSSARIO DELLA MADONNA DELLA SACRA LETTERA: PROSPETTIVE D'INDAGINE

Nel 1712, Giovanni Calì – acese – fonda la chiesa della Madonna della Sacra Lettera. I lavori di costruzioni si ultimeranno nel 1761, e da quel momento, in questo luogo sacro troveranno sepoltura i "coloni" ripostesi [42, 43]. E anche se vero che la borgata è dentro l'amministrazione mascalese, la comunità marinara è già presente e stanziata in quegli spazi, testimoniando il fatto che le attività commerciali si spostavano dal centro della Contea verso la "periferia" costituita dal porto naturale.

Il secolo successivo, Riposto otterrà l'autonomia, la qual cosa sarà determinante per il destino del sepolcro, che dovette essere chiuso – nessuno riuscì più ad accedervi –, in conformità con la normativa sanitaria di polizia funeraria, la quale, proprio in quel periodo, aveva definito un nuovo approccio politico, culturale e giuridico con la morte e le sepolture.

Nel corso dell'ultimo cinquantennio, il luogo ha attirato l'attenzione di diversi studiosi di fama nazionale ed internazionale, che sentivano la necessità di scrivere la storia della Città di Riposto – come Giuseppe Giarrizzo, storico e accademico dei Lincei, la quale si è sempre mostrata dimensione territoriale di spiccata importanza politica e commerciale per tutto l'hinterland.

Alla visione storico-culturale, si accosta quella prettamente urbanistica: la chiesetta ha importanti problemi strutturali che richiedono interventi manutentivi. Sicché, negli anni Settanta del Novecento, si cercò di risolvere il problema delle infiltrazioni, dovute alla vicinanza del mare. E proprio nel 1979, durante gli interventi di sventramento, sorse il sospetto che al di sotto del calpestìo della chiesetta potessero trovarsi tracce utili allo studio della sua struttura e delle sue origini.

Venne rimossa, così, parte della massicciata in cemento; dapprima ci fu l'impatto con un banco di ghiaia marina mista a sabbia, dell'altezza di circa 50 cm, formatosi probabilmente a causa di un'improvvisa e violenta mareggiata. Successivamente, si venne a contatto con un piano pavimentato che mostrava una posa di piastrelle in cotto di forma quadrata. Ancora oltre, si arrivò al rinvenimento di una non trascurabile quantità di ossa umane.

L'intervento della Sovrintendenza ha fatto sì che le ossa venissero racchiuse, tutte insieme, in spazi protetti da vetri, al fine di consentire l'ingresso ad eventuali visitatori e studiosi.

Oggi, in una visione di sviluppo culturale digitale dei luoghi – nel momento storico in cui le metodologie di ricerca possono avvalersi di strumenti tecnologici in grado di rispondere con efficacia alla creazione di una società "*data intensive*" –, il progetto "We Are What They Were" guarda alla possibilità di procedere ad uno studio che possa muoversi su più livelli. Da un lato, restituire valore storico, culturale e metodologico ad un luogo che presenta delle caratteristiche uniche, al fine di soddisfare l'incessante richiesta di costruzione di *prodotti digitali* che possano immettere nella rete internet il patrimonio storico "analogico", e, da un altro lato, effettuare una serie di studi genetici [7, 9, 13, 14, 20, 21, 27, 33] in grado di collegare i viventi ripostesi all'antenato defunto, le cui spoglie hanno trovato riposo, nel 18° secolo, nella cripta del citato luogo sacro.

L'analisi del DNA dei teschi (vd. Figg. 6, 7), consentirebbe, infatti, di definire scientificamente la struttura della comunità, lo stile di vita alimentare, il quadro epidemiologico, nella prospettiva di creare un complesso di dati interconnessi (Big Data) che possa essere esplicazione di un *workflow* metodologico in grado di spiegare l'uomo nella sua essenza unica ed irripetibile.

I «resti umani» sono tutto ciò che il tempo conserva degli individui, delle popolazioni o delle specie del Passato. Oggi, queste "identità" diventano concetti centrali della ricerca scientifica, in ogni ambito, rappresentando, però, una "questione" difficile da trattare, sia per le problematiche interpretative, ma, ancora di più, per la dimensione etica che emerge con forte preponderanza. I resti umani rappresentano una singolarità che è unica, più che rara: sono – insieme – il "soggetto" che è stato in vita, ma anche l'"oggetto", in quanto resti materiali, di studio dell'Antropologia. Si pongono, dunque, in una sorta di drammatica condizione intermedia tra quel che rimane di un'esistenza e, al contempo, ciò che rappresenta un'insostituibile testimonianza d'interesse scientifico, un vero e proprio *archivio biologico* e culturale degli esseri umani del passato, che può (e deve) essere registrato, conosciuto e interpretato. Essi sono *fonte* di informazioni uniche circa l'evoluzione e la filogenesi umana, l'ecologia delle popolazioni e consentono di descrivere le dinamiche migratorie. I resti umani sono essenziali per ricostruire – ad esempio – lo

stile e la qualità di vita nel passato (es. comportamenti alimentari, funerari, pratiche mediche e terapeutiche, attività svolte in vita) come anche aspetti paleoepidemiologici nello studio delle malattie, e diventano tanto più rilevanti quanto più scarsi sono altri tipi di documentazione.

Appare chiaro, quindi, che sia necessaria una profonda riflessione, ma, soprattutto, aprire la frontiera della ricerca verso quella interdisciplinarità e multidisciplinarità in grado di arricchire la conoscenza dell'Uomo e del Mondo – anche digitale. Per questo motivo, è necessaria una presa di coscienza scientifica che porti l'attenzione degli studiosi su uno degli ossari siciliani più corposi dell'isola: la cripta della Chiesa della Madonna della Sacra Lettera di Riposto.

Figura 6

Figura 7

BIBLIOGRAFIA

[1] Amico, Vito Maria. *Lexicon Topographicum Siculum*. Palermo, 1757.

[2] Barbera, Giuseppe. *Il giardino del Mediterraneo: Storie e paesaggi da Omero all'Antropocene*. Milano: Il Saggiatore, 2021.

[3] Bevilacqua, Piero. *Demetra e Clio: uomini e ambiente nella storia*. Roma: Donzelli, 2001.

[4] Bevilacqua, Piero. *La terra è finita: breve storia dell'ambiente*. Bari: Laterza, 2014.

[5] Bleichmar, Daniela, e Vanessa R. Schwartz. «Visual History. The Past in Pictures». *Representations* 145 (2019): 1–31.

[6] Botstein, David, Raymond L. White, Mark Skolnick, e Ronald W. Davis. «Construction of a genetic linkage map in man using restriction fragment length polymorphisms». *American Journal of Human Genetics* 32 (1980): 314–31.

[7] Bowcock, A.M., C. Bucci, J.M. Hebert, JR. Kidd, K.K. Kidd, J.S. Friedlaender, e Luigi Luca Cavalli-Sforza. «Study of 47 DNA markers in five populations from four continents». *Gene geography : a computerized bulletin on human gene frequencies,* 1 (1987): 47–64.

[8] Bradley, Adam James, Mennatallah El-Assady, Katharine Coles, Eric Alexander, Min Chen, Christopher Collins, Stefan Janicke, e David Joseph Wrisley. «Visualization and the Digital Humanities». *IEEE Computer Graphics and Applications* 38 (2018): 26–38.

[9] Brandt, Guido, Wolfgang Haak, Christina J. Adler, Christina Roth, Anna Szécsényi-Nagy, Sarah Karimnia, Sabine Möller-Rieker, et al. «Ancient DNA reveals key stages in the formation of Central European mitochondrial genetic diversity». *Science* 342 (2013): 257–61.

[10] Braudel, Fernand. *Civiltà e imperi del Mediterraneo nell'età di Filippo II*. Milano: Mondadori, 2011.

[11] Braudel, Fernand. *Il Mediterraneo, lo spazio, la storia, gli uomini, le tradizioni*. Milano: Bompiani, 1992.

[12] Braudel, Fernand. *Scritti sulla storia*. Milano: Bompiani, 2003.

[13] Cann, Rebecca L., Mark Stoneking, e Allan C. Wilson. «Mitochondrial DNA and human evolution». *Nature* 325 (1987): 31–36.

[14] Caramelli, David, e Martina Lari. *Il DNA antico: Metodi di analisi e applicazioni*. Firenze: A. Pontecorboli, 2004.

[15] Cavalli-Sforza, L.L., e A.W.F. Edwards. «Phylogenetic analysis: Models and estimation procedures». *Evolution Evolution* 21 (1967): 550–70.

[16] Cavalli-Sforza, L.L., J.R. Kidd, K.K. Kidd, C. Bucci, A.M. Bowcock, B.S. Hewlett, e J.S. Friedlaender. «DNA Markers and Genetic Variation in the Human Species». *Cold Spring Harbor Symposia on Quantitative Biology Cold Spring Harbor Symposia on Quantitative Biology* 51 (1986): 411–17.

[17] Cavalli-Sforza, Luigi Luca. «Genes, peoples, and languages». In *Proceedings of the National Academy of Sciences of the United States of America*, 94:7719–24. Harvard Magazine, 1997.

[18] Cavalli-Sforza, Luigi Luca. *Geni, popoli e lingue*. Milano: Adelphi, 1996.

[19] Cavalli-Sforza, Luigi Luca. *L'evoluzione della Cultura*. Torino: Codice, 2004.

[20] Cavalli-Sforza, Luigi Luca. «Population structure and human evolution». In *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character*, 362–79, 1966.

[21] Cilli, Elisabetta, e Mirko Traversari. *Le Mummie di Roccapelago. Un Progetto Pilota Di Ricerca Interdisciplinare Tra Archeologia, Antropologia, Storia e Scienze Applicate*. Bologna: Istituto per i Beni Artistici, culturali e naturali Regione Emilia-Romagna, 2020.

[22] Cohen, Daniel J., e Roy Rosenzweig. *Digital History, a Guide to Gathering, Preserving, and Presenting the Past on the Web*. Philadelphia: University of Pennsylvania Press, 2006.

[23] De Lorenzo, Renata. *Storia e misura: indicatori sociali ed economici nel Mezzogiorno d'Italia, secoli XVIII-XX*. Milano: FrancoAngeli, 2007.

[24] Di Marzio, Giacchino. *Dizionario topografico della Sicilia di Vito Maria Amico tradotto dal Latino e continuato sino ai nostri giorni*. tipografia di Pietro Morvillo, 1856.

[25] Ebbrecht-Hartmann, Tobias, Noga Stiassny, e Lital Henig. «Digital visual history: historiographic curation using digital technologies». *Rethinking History* 27 (2023): 159–86.

[26] Ercole, Francesco. *I riveli di beni e di anime del Regno di Sicilia*. Istituto Poligrafico dello Stato, 1931.

[27] Francalacci, Paolo, Giovanna Melas, e Domenica A. Obinu. «Estrazione e analisi del DNA da reperti museali». In *Atti dei seminari ANMS di Pavia*, 24–30, 2008.

[28] Giarrizzo, Giuseppe. *La Sicilia dei terremoti. Lunga durata e dinamiche sociali*. Catania: Maimone, 1997.

[29] Kan, Y.W., e A.M. Dozy. «Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation». In *Proceedings of the National Academy of Sciences of the United States of America*, 75:5631–35, 1978.

[30] Kenneally, Christine. *Storia invisibile della razza umana: Come il DNA e la storia danno forma alla nostra identità e al nostro futuro*. Milano: Mondadori, 2016.

[31] Krings, M., A. Stone, R.W. Schmitz, H. Krainitzki, Mark Stoneking, e Svante Pääbo. «Neandertal DNA sequences and the Origin of Modem Humans». *Celi*, 1997, 19–30.

[32] Ligresti, Domenico. *Dinamiche demografiche nella Sicilia moderna: 1505-1806*. Milano: FrancoAngeli, 2002.

[33] Manzi, Giorgio, Maria Giovanna Belcastro, e Jacopo Moggi Cecchi. *Quel che resta. Scheletri e altri resti umani come beni culturali*. Bologna: Il Mulino, 2022.

[34] Padoa-Schioppa, Emilio. *Antropocene. Una nuova epoca per la Terra, una sfida per l'umanità*. Bologna: Il Mulino, 2021.

[35] Padoa-Schioppa, Emilio. *Storia ecologica dell'Europa. Un continente nell'Antropocene*. Bologna: Il Mulino, 2023.

[36] Rosenzweig, Roy. «The Road to Xanadu: Public and Private Pathways on the History Web». *Journal of American History* 88 (2001): 548–79.

[37] Scalisi, Lina. *Catania. L'identità urbana dall'antichità al Settecento*. Catania: Sanfilippo, 2009.

[38] Scalisi, Lina. *Potere e sentimento*. Roma: Edizioni di Storia e Letteratura, 2023.

[39] Shaw, Jonathan. «Who killed the men of England? The written record of history meets genomics, evolution, demography, and molecular archaeology». *Harvard Magazine*, 2009, 30-35/75. https://www.harvardmagazine.com/sites/default/files/pdf/2009/07-pdfs/0709-30.pdf.

[40] Sorre, Maximilien-Joseph. *Les fondements biologiques de la géographie humaine*. Colin, 1943.

[41] Spina, Salvatore. *Archivio storico del comune di Riposto: Inventario*. Catania: Maimone, 2013.

[42] Spina, Salvatore. *Riposto. Territorio, infrastrutture, identità urbana, 1841-1920*. Viagrande: Algra, 2015.

[43] Spina, Salvatore. *Riposto vecchio e Riposto nuovo negli atti notarili di Giovanni Calì e Geronimo Pasini: studi per la Storia di Riposto*. Acireale-Roma: Bonanno Editore, 2011.

[44] Spina, Salvatore, e Giuseppe Sorbello. «Dagli archivi storici alle mappe genomiche. Il caso di Riposto». *Aidainformazioni* 1, fasc. 2 (2021).

[45] Theibault, John. «Visualizations and Historical Arguments (Theibault)». In *Writing History in the Digital Age*, a cura di Jack Dougherty e Kristen Nawrotzki. Michigan: University of Michigan Press, 2012.

[46] Vitali, Stefano. *Passato digitale, le fonti dello storico nell'era del computer*. Milano: Mondadori, 2004.

[47] Wainscoat, J.S., A.V.S. Hill, A.L. Boyce, J. Flint, Hernandez, Thein S.L., J.M. Old, et al. «Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms» *Nature* 319 (1986): 6–12.

[48] Zangheri, Renato. *Catasti e storia della proprietà terriera*. Torino: Einaudi, 1980.

Drawing history from the Codice Pelavicino documents: Graph Visualization for Human Researchers

Natthida Wiwatwicha
University of Pisa, Italy – n.wiwatwicha@studenti.unipi.it

ABSTRACT

This paper reports the lessons from development of tools to visualize relationships between entities in the Codice Pelavicino, which already exists in a Digital Edition. The objective of the visualization system is to facilitate the exploration and reading of the Codice Pelavicino for historical research, while supporting changes in interpretative layers and multiple research inquiries. This project investigates graph representation techniques, with focus on legibility and usability issues because research processes around medieval documents and digital history heavily involve human consultations and collaboration. This paper demonstrates the current version of the relationship graphs between the Bishop of Luni and various groups of entities over two centuries.

KEYWORDS

Graph visualization; Digital Humanities; Digital History; Human-centered research.

1. INTRODUCTION

Navigating historical documents can be challenging, despite the availability of digital scholarly editions. Network graph visualization is suitable for representing relationships between entities in semi-structured data, such as XML formats. However, network graphs present usability challenges [1], and as visual artifacts, they often merely accompany findings rather than contributing to the process of obtaining them [2]. Network graphs also present methodological issues: tools for graph analysis are often opaque to digital humanities researchers, and the task of graph visualization can devolve into merely "drawing complicated graphs" [4: 23-25]. Nevertheless, numerous efforts continue to utilize network graphs for text representation to facilitate exploration and research [3, 9][1]. Through iterative prototyping, this case study develops techniques for graph visualization that convey relationships between entities in the XML dataset of the Codice Pelavicino Digital Edition project [8]. The ongoing nature of data interpretation and research inquiries provides a unique opportunity to understand and develop network visualizations with an emphasis on versatility and sustainability. This poster contribution discusses the procedures and decisions involved in creating graph visualizations in Neo4j from an XML dataset, and evaluates the usefulness of the proposed approach.

2. PRECEDENTS AND PRELIMINARIES

Recent projects that highlight the use of graph visualization as a tool for exploration of documents generate small, sparse graphs with few nodes, enabling researchers to independently identify and analyze patterns without relying on pre-existing algorithms. The Leopardi's Zibaldone project transforms an XML dataset into an explorable semantic network, where text fragments are depicted as nodes [9]. In a knowledge management context, Zabir Said's graph explorer offers users complete freedom to filter entity types[2]. Intergraph for historical documents allows users to view named entities and their co-appearance relations in multiple small subgraphs, facilitating easier navigation and comparison over time [3]. The simplicity and generality of the graph data model allow for the expansion of datasets to include additional documents from new sources [3].

Although graph visualization projects often start after the completion of dataset which simplifies and trivializes data management, some researchers still prefer to use graph databases over spreadsheets or relational database for their efficiency, simplicity, integrity, and scalability [3, 5, 9]. Graph database management systems (GDBMSs) are favored in various fields for their ability to integrate diverse and loosely structured data efficiently, while relational models can become overly complicated [5] or takes more computational resources [9].

The Codice Pelavicino (CP) Digital Edition's dataset introduces unique challenges for graph visualization. The visualization must meet the needs of medieval historians in supporting questions characterized by time periods and various

[1] See Khoo, Christopher. "Explore a Zabir Said Knowledge Graph". Zubir Said Knowledge Graph. July 2021. https://zubirsaid.sg/ZS.graph.html

[2] *Ibidem.*

objects in the codex, while also enabling the direct extraction and contextual viewing of data [7]. As an ongoing, collaborative project with the XML tagging still in progress in a semi-automatic and manual manner, the CP Digital Edition prioritizes accessibility, transparent editing, and incremental releases which require the visualization system to support rapid deployment and review for validation. The population of XML tags and the graph's specifications are determined by such use case and the domain knowledge of experts contribute to the DE. The use of ready-made solutions for creating network graphs and extracting data from XML documents can be costly to validate and modify to fulfill these requirements.

3. PROTOTYPING NETWORK GRAPHS FROM THE CODICE PELAVICINO DIGITAL EDITION

The project employs a process-focused design research method, adopting an iterative approach, with cycles of rapid prototyping for data extraction, processing, and visualization design. Each step is validated for correctness and concludes with user evaluation. Visualization development involves design exploration and experimentation to compare various methods. Competing schemes are developed concurrently until user interviews reveal a definitive best design. For data extraction and processing, existing XML tags such as “persName,” “roleName,” and “orgName” are used for named-entity extraction. Any inconsistencies not arising from the extraction and processing scripts are documented and reported upstream to pinpoint potential errors.

Entity – Entity Schema vs. Entity – Document Schema

Creating a database necessitates defining a data schema before its creation, even before XML tags are finalized or its use case is fully specified; a graph database simplifies this process. Representing relationships between entities through co-occurrences at the document level supports these conditions. The only required bi-directional relation to capture is that an entity "is present in" a unit of text.

Relationship graphs typically connect entities. For example, social network graphs present people as nodes. However, the entity-entity schema faces a critical issue: its ambiguity in depicting relationships involving more than two entities. Figure 1 illustrates that the same representation could imply that three entities either have some connection as a group or are related merely in pairs, without any group interaction. This schema fails to differentiate these scenarios.

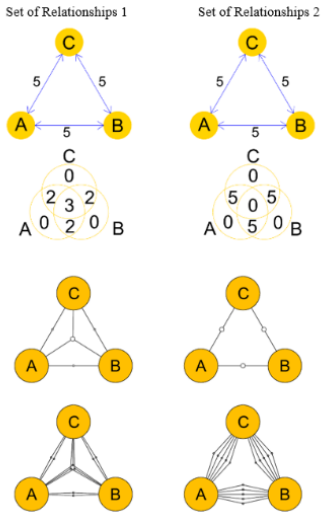


Figure 1. Two identical entity-entity graphs (row 1) are generated from two different set relationships (row 2) because the schema cannot capture the interactions that include three entities (A, B, and C) simultaneously. To resolve this issue, intermediary nodes can be introduced to represent the intersection of three sets, using a single node to store a value (row 3) for each relationship, or a single node to represent each instance (row 4).

The Entity-Document schema facilitates the identification of co-occurrences involving more than two entities and allows for the storage of more document attributes in the database, rather than as attributes of relationships. Figure 2 demonstrates how users can inspect documents that are sites of co-occurrence between entities. Additionally, this schema supports queries and analyses using a distance parameter similar to those in the Entity-Entity graph, with minor adjustments—for instance, the distance or degree of separation between entities is effectively doubled. Therefore, the Entity-Document schema is preferred in this case study (see Fig.1, bottom row).

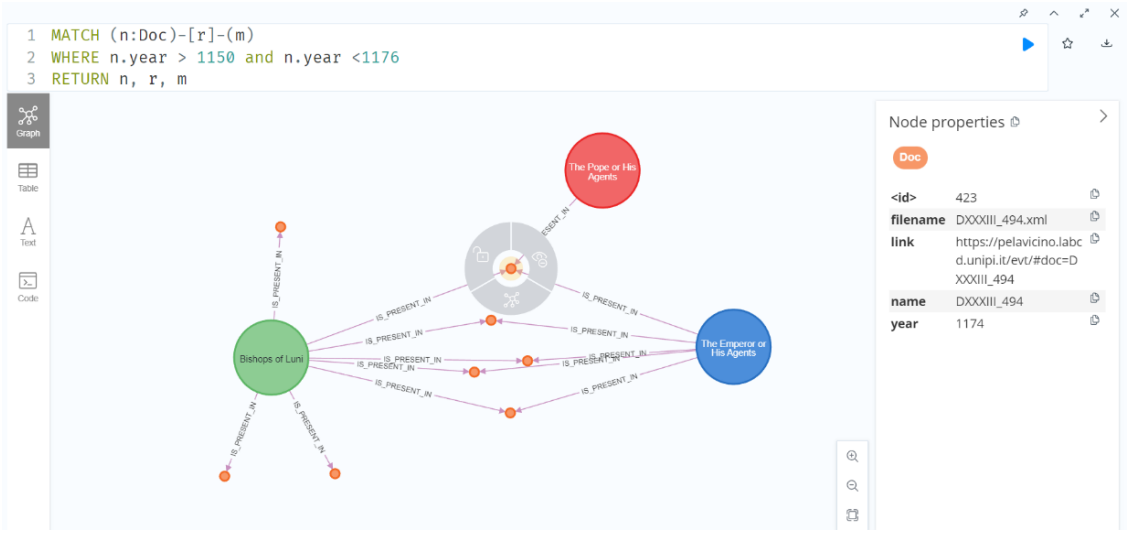


Figure 2. A relationship graph between the Bishop of Luni (green), the Pope (red), and the Emperor (blue) between 1150-1176 in Neo4j Browser. The document nodes reveal the document ID, year, filename, and the link to the document in the Codice Pelavicino Digital Edition.

Text extraction and management

To produce graphs from XML documents, specific functions are developed for each task:

Extracting Entities: A simple Python script uses the built-in ElementTree (ET) with an lxml parser to extract predictable entities such as <doc> and their attributes. For named entities with irregular tag structures (e.g., persons, organizations, roles, places), Beautiful Soup (BS) is employed due to its capability to handle complex tags. Comparing results from ET and BS aids in understanding the data. Future plans include testing ready-made named-entity recognition solutions for medieval Latin as a benchmark.

Aggregating Entities: Aggregation occurs during data pre-processing and is straightforward in Python using rule-based scripts. Simple tasks, such as grouping organizations as civic or religious, or more complex tasks involving .csv input for categorizing job types, require consultation and validation with human experts with varying levels of involvement. The graph database management system Neo4j and its visualization tools include a library (APOC) capable of creating rule-based virtual nodes, but this feature is not yet used at this prototyping stage due to its unpredictable effects on the user interface of the visualization.

Converting Data for Graph Database: A simple script converts entities, synthetic (aggregated) entities, and documents into nodes, and their occurrences in documents into relationships. This data is formatted into input types supported by the Neo4j graph database (e.g. .json, .csv, or Cypher code), based on the graph's schema.

Once the data is imported into the graph database, interactive graph visualizations become instantly available. Users can query specific time periods and filter entities in the Neo4j Browser or in Neo4j Bloom via a GUI slicer feature, allowing them to study the changes in the graph over time by leveraging its automated histogram.

4. EVALUATION OF GRAPHS

Enhancing graph legibility and clarity involves more than just schema selection; it requires thoughtful data preparation and strategic visual customization to effectively support user exploration and understanding. Graph legibility is greatly enhanced by text management: categorizing entities to merge them into virtual nodes, rather than only filtering them, which alone is insufficient for clarifying information. Users find that including metadata about the logic of their aggregation rules in the synthetic node and linking to related sources aids verification process. Limiting each study to a single theme or variable also reduces the number of elements. Figure 3 displays the graph before entity categorization and scope restriction, whereas Figure 4 shows the simplified graphs useful for studying a set of relationships through time, or within the same period across subjects of study.

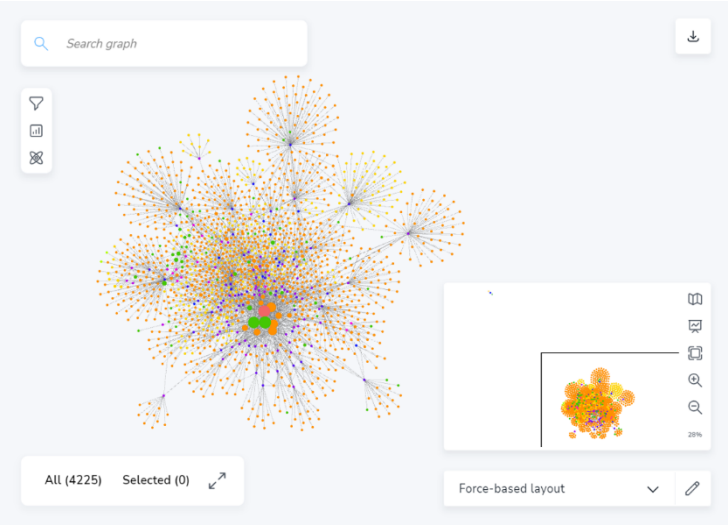


Figure 3. An earlier prototype from sub-sample extracted from the Codice Pelavicino XML dataset, showing individuals (orange) and institutions (green) as they relate to one another through the documents (blue), without any categorization of entities to aggregate the nodes or restrict the scope of study. The Bishop of Luni (red) is seen in the center, and surrounded by a few institutions and individuals identified as notaries by their XML tag.

Additionally, offering more querying possibilities can complicate the extraction of useful information from the graph. Users often struggle to know precisely what to look for, and even knowledgeable users may not understand whether their queries will yield meaningful results. This observation supports a hypothesis that experts familiar with the digital edition's structure, nature of dataset, and contextual information should selectively enable the most promising search functions [7].
In terms of visual detail, experimentation and user feedback indicate that increased freedom in visual attribute customization does not always aid graph interpretation. For instance, representing a parameter by node size is redundant if it is already indicated by the number of edges. Conversely, removing unnecessary entities is not always beneficial; maintaining visibility of all 529 document nodes, regardless of their connections, provides context about the proportion of occurrences in-focus relative to total occurrences.

Finally, user interviews reveal that while graphs are valuable for identifying initial research directions and detecting patterns, a single graph visualization cannot encompass the exploration process of an entire set of documents nor yield standalone conclusions. Graphs help illustrate relational dynamics and is useful for identifying intensities across time periods, but these findings must be integrated with additional investigations.

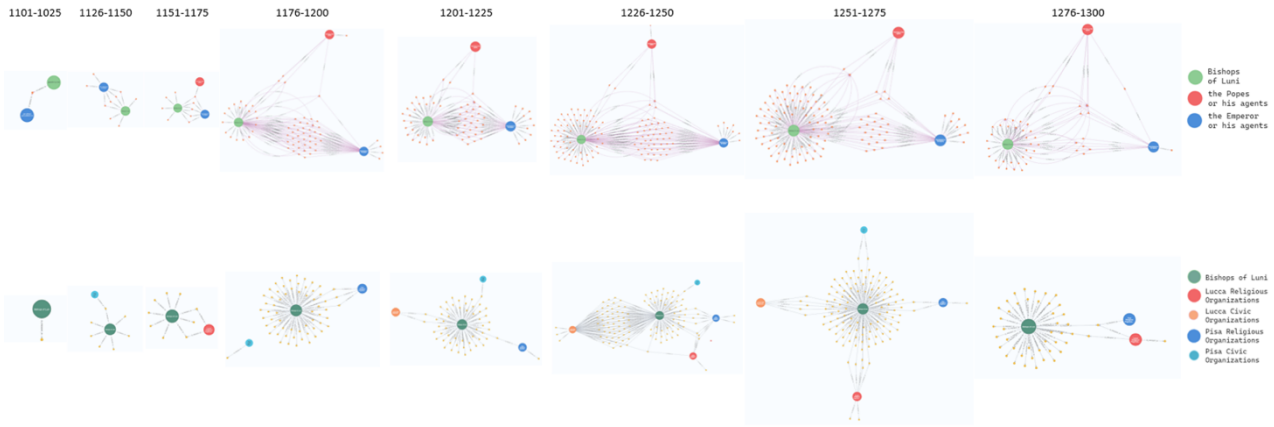


Figure 4. Snapshots of graphs representing relationships by co-occurrence in the codex at 25-year intervals from 1100 to 1300. The top row illustrates the evolving relationships among the Bishop of Luni (green), the Pope (red), and the Emperor (blue). The bottom row displays the interactions between the Bishop of Luni and the religious/civic organizations of Pisa (blue/teal) and Lucca (red/orange). In the Bishop of Luni-Pope-Emperor sequence, the curved edges represent multiple bishops' involvement in a document.

5. GENERATING VISUALIZATION SCENES FOR MULTIPLE RESEARCH QUESTIONS AND PURPOSES

Using the presented process, multiple studies from the Codice Pelavicino, such as changes in relationships between the

bishop of Luni, the Pope, and the Emperor, as well as between the bishop and various religious and civic institutions organized by place, can be visualized (see Fig. 4). Another ongoing study examines the relationships between the bishop and individuals of different occupations. These studies required only minor adjustments in the extraction rules, while the rest of the workflow remained consistent. Additionally, with these tools and entity-document schema, new areas for visualization and study are accessible. For example, by introducing a new relationship type "is member of," it is possible to visualize the presence or absence of organizational affiliations of the same person in different documents. A future relationship type "is associated with" could link new node types such as categories, places, or concepts, and allow for the addition of metadata to those nodes. The same scripts, with minor modifications to the data converter function, can also support an entity-entity graph. In summary, the data preparation system along with graph database enables the visualization and analysis of various relationships within the Codice Pelavicino, with flexibility to adapt and expand studies through minor modifications.

6. CONCLUSION

Transferring a manuscript to the digital space opens up possibilities of various formats and accompanying tools. Network graphs can support Digital Editions to overcome their limitations from the "paginated form" [6], increases its browsability, and facilitates its reading. Graphs generated from an XML dataset allow researchers to navigate hypertext differently, finely control levels of contextual information, and locate necessary sections more efficiently. The document-entity graph schema offers more accurate cooccurrence representation compared to the entity-entity schema, and provides more information and shortcuts to aid document inspection in validating interpretations and edits of the original data. Keeping graphs small and straightforward as "graphlets" [4: 31] is one possible technique to help digital historians and humanists reimagine the contributions of digital tools to support their research. The design decisions presented here aim at creating more legible, verifiable, and versatile graph visualizations, and improving accessibility to textual data sources by bridging distant-reading and close-reading. It is hoped that this study will be beneficial to others interested in drawing network graphs for exploring and navigating texts and archives.

7. ACKNOWLEDGEMENTS

This contribution received support from the Laboratory of Digital Cultures at the University of Pisa and the Ministry of International Cooperation and Foreign Affairs, the dataset from The Codice Pelavicino Digital Edition Project, and guidance from professor Enrica Salvatori and professor Vittore Casarosa.

REFERENCES

[1] AlKadi, Mashael, Vanessa Serrano, James Scott-Brown, Catherine Plaisant, Jean-Daniel Fekete, Uta Hinrichs, and Benjamin Bach. ‘Understanding Barriers to Network Exploration With Visualization: A Report from the Trenches’. *IEEE Transactions on Visualization and Computer Graphics*, 2022. https://doi.org/10.1109/tvcg.2022.3209487.

[2] Bach, Benjamin, Nathalie Henry Riche, Roland Fernandez, Emmanoulis Giannisakis, Bongshin Lee, and Jean-Daniel Fekete. ‘NetworkCube: Bringing Dynamic Network Visualizations to Domain Scientists’. In *Conference on Information Visualization (InfoVis)*. Chicago, United States, 2015.

[3] Bornhofen, Stefan, and Marten Düring. ‘Exploring Dynamic Multilayer Graphs for Digital Humanities’. *Applied Network Science* 5, no. 1 (2020). https://doi.org/10.1007/s41109-020-00295-x.

[4] Fiscarelli, Antonio Maria. ‘Social Network Analysis for Digital Humanities’. In *Digital History and Hermeneutics: Between Theory and Practice*, edited by Andreas Fickers and Juliane Tatarinov, 23–42. Berlin, Boston: De Gruyter Oldenbourg, 2022. https://doi.org/10.1515/9783110723991-002.

[5] Healy, Meadhbh, Thomas O’Connor, and John Keating. ‘Comparison of Graph- and Collection-Based Representations of Early Modern Biographical Archives’. In *Graph Technologies in the Humanities - Proceedings 2020*, edited by Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera, and Joris van Zundert, 2020. http://ceur-ws.org/Vol-3110/paper4.pdf.

[6] Sahle, Patrick. ‘What Is a Scholarly Digital Edition?’ In *Digital Scholarly Editing: Theories and Practices*, edited by Elena Pierazzo and Matthew J. Driscoll, 1–22. Cambridge: Open Book Publishers, 2016. https://doi.org/10.11647/OBP.0095.

[7] Salvatori, Enrica. ‘Appetite Comes With Eating. The Never Ending Digital Edition of the Codex Pelavicino’. *Umanistica Digitale*, no. 10 (2021). https://doi.org/10.6092/ISSN.2532-8816/12577.

[8] Salvatori, Enrica, Edilio Riccardini, Roberto Rosselli del Turco, Laura Balletto, Chiara Alzetta, Chiara Di Pietro, Chiara Mannari, Raffaele Masotti, and Alessio Miaschi. *Codice Pelavicino. Edizione digitale*, 2020. https://doi.org/10.13131/978-88-944430-2-8.

[9] Stoyanova, Silvia. ‘Articulating Intra- and Intertextual Relationships in the Fragment Collection. Working with the Digital Edition of Giacomo Leopardi’s Zibaldone’. *Magazén* 4, no. 1 (2023): 13–42. https://doi.org/10.30687/mag/2724-3923/2023/01/001.

[10] Timón-Reina, S., M. Rincón, and R. Martínez-Tomás. ‘An Overview of Graph Databases and Their Applications in the Biomedical Domain’. *Database (Oxford)*, 2021. https://doi.org/10.1093/database/baab026.

Gestione informatica della documentazione archeologica "minore". Metodologie e applicazioni nell'ambito del progetto Storage

Marianna Figuera[1], Erica Platania[2]
[1]Dipartimento di Scienze Umanistiche, Università di Catania, Italia - marianna.figuera@unict.it
[2]Dipartimento di Scienze Umanistiche, Università di Catania, Italia - erica.platania@unict.it

ABSTRACT[1]

Lo sviluppo della *Digital Archaeology* ha permesso di affrontare da nuove prospettive le tematiche della gestione, archiviazione e condivisione dei dati archeologici. Nel presente contributo si affronta il tema dell'uso di strumenti di gestione informatica per la ricerca archeologica, partendo da due casi studio sull'analisi degli *small finds* da due siti minoici cretesi e sui fenomeni pastorali della Sicilia orientale in età preistorica. Entrambi trattano classi di reperti a lungo sottostimate che, in realtà, dispongono di un potenziale informativo notevole, la cui gestione digitale implica problematiche di ordine metodologico di non immediata risoluzione. Nell'ambito del progetto interdipartimentale "Storage. Dai dati al Web" è stato sviluppato, sulla base di due precedenti lavori di ricerca, un sistema di gestione dei dati che potesse rispondere alle criticità riscontrate, permettendo un avanzamento della ricerca nei due rispettivi campi di studio. Il database relazionale per il trattamento dei reperti archeologici cd. "minori" è stato progettato in funzione della conservazione della integrità, provenienza, trasparenza e riproducibilità dei dati, evitando ogni sorta di semplificazione e rendendo il sistema un "contenitore di memorie".

PAROLE CHIAVE

Small Finds; resti faunistici; *Legacy Data*; memoria; Patrimonio Digitale.

1. INTRODUZIONE

Il rapporto tra Archeologia e Informatica non è scevro da problematiche di natura metodologica, insite fondamentalmente nella mancanza di un linguaggio di base comune alle due discipline. L'incertezza del dato archeologico, la lettura soggettiva del materiale di scavo e la sua interpretazione variabile nel tempo, hanno condotto a ritenere gli strumenti informatici spesso inadeguati al trattamento e alla risoluzione delle problematiche archeologiche [3]. Fra gli aspetti più delicati dell'interazione tra le due discipline vi è il riuscire a preservare la complessità del modo di operare dell'archeologo e la possibilità o meno di adottare un linguaggio rigoroso nel trattare argomenti di natura archeologica [7: 21]. Un aspetto che riguarda trasversalmente tutti i dati archeologici è anche quello della gestione dei cosiddetti *legacy data*: termine, nato nel settore informatico, e che nell'ambito della *Digital Archaeology* definisce la grande quantità di informazioni accumulata negli archivi, o che continua ad essere prodotta secondo metodi e procedure tradizionali, e che quindi va digitalizzata o, se già digitalizzata, può risultare inadeguata per un trattamento informatico aderente agli attuali standard. L'uso del termine *legacy* non è necessariamente legato al concetto di obsoleto, ma si riferisce appunto ai dati – particolarmente abbondanti in archeologia – che prima di poter essere usati in ambiente digitale devono essere "preparati e manipolati" [1]. Dal punto di vista dell'informatico, inoltre, la gestione del dato archeologico è spesso problematica per la difficoltà di ottenere una informazione non inficiata dal processo interpretativo. Infatti, è di fondamentale importanza accettare il fatto che nella natura stessa del dato archeologico è insita spesso l'indeterminatezza: le imprecisioni possono riguardare una serie di aspetti, fra cui quello cronologico, spaziale, funzionale, ecc. L'atto classificatorio, inoltre, è fortemente influenzato da variabili in parte razionali e in parte intuitive, dal processo interpretativo e, infine, dall'esperienza stessa del ricercatore, pertanto può essere parzialmente soggettivo [13: 281]. La gestione informatica del dato archeologico finisce, quindi, per ottimizzarne i processi di acquisizione e trattamento, spingendo gli archeologi ad effettuare controlli maggiori sulla qualità e l'organizzazione dei dati, per meglio rispondere ai bisogni degli addetti ai lavori e degli utenti. In particolare, uno degli obiettivi primari da perseguire è quello del perfezionamento degli aspetti di documentazione e di riuso, applicando criteri quantitativi e qualitativi che coinvolgano i concetti di integrità, provenienza, trasparenza e riproducibilità dei dati, investendo anche l'aspetto dell'interpretazione [6: 53-54].

[1] Marianna Figuera è responsabile del caso studio degli *small finds*; Erica Platania è responsabile del caso studio sui fenomeni pastorali nella Sicilia Preistorica. La scrittura del contributo è di responsabilità di entrambe le a., tranne dove espressamente indicato.

2. CARATTERISTICHE E PROBLEMATICHE DELLA DOCUMENTAZIONE ARCHEOLOGICA "MINORE"

Problemi metodologici legati al trattamento dei dati archeologici emergono, in particolar modo, quando ci si confronta con quelle categorie di reperti considerati tradizionalmente come "minori". Fra questi vi sono *in primis* i cosiddetti *small finds*, termine con cui vengono comunemente definiti reperti realizzati in diversi materiali che non rientrano nelle grandi categorie dei contenitori vascolari e delle produzioni suntuarie storico-artistiche [6: 25-30]. È possibile considerare come *small finds* manufatti connessi con un'ampia sfera di aspetti della vita quotidiana: coinvolti in molteplici attività domestiche, in azioni di tipo liturgico, rituale, di carattere militare e, soprattutto, di carattere artigianale e produttivo.

Queste categorie di reperti hanno caratteristiche tali da porre una serie di problematiche metodologiche relative alla documentazione e alla gestione informatica dei loro dati. Si tratta, infatti, di reperti solitamente non diagnostici dal punto di vista cronologico e funzionale, che si caratterizzano per la polifunzionalità, ovvero la possibilità che una stessa tipologia di manufatti possa aver svolto compiti diversi. Di conseguenza, il loro potenziale informativo è stato spesso sottostimato, rimanendo ai margini degli interessi scientifici degli archeologi, trattati marginalmente nei lavori dedicati allo studio dei materiali dove, per consuetudine, è stata riservata a questi "*residual finds*" una sezione posta in coda [12: 79]. In realtà, in quanto prodotti della cultura materiale sono non "oggetti" passivi ma "soggetti" che hanno avuto parte attiva nei processi di interazione sociale, quindi le loro potenzialità interpretative emergono una volta compresi i contesti o gli ambiti sociali di appartenenza [20:190, 2: 66-67][2].

Un'altra tipologia di reperti a lungo considerata minore è costituita dai resti faunistici, che rappresentano una fonte preziosa di informazioni per lo studio del rapporto uomo-animale nel tempo. Le ossa provenienti dai siti archeologici possono fornire informazioni non solo sulla dieta, l'igiene, il clima, la stagione di occupazione di un sito, i metodi di caccia e allevamento, la pastorizia, ma contribuire, talvolta in maniera dirimente, allo studio di problematiche di più ampio respiro come il commercio, la produzione di strumenti, la religione e il rituale funebre [4: 3]. Tuttavia, collocandosi al confine tra discipline umanistiche e scienze biologiche, lo studio delle ossa animali da contesti di scavo è stato a lungo considerato sussidiario e perciò non pienamente incorporato nella interpretazione archeologica dei contesti, sottostimandone il potenziale informativo, con conseguenze rilevanti dal punto di vista della condivisione, diffusione e conservazione dei dati prodotti. Nell'ultimo cinquantennio, l'avanzamento metodologico e l'applicazione di tecnologie digitali all'ambito archeozoologico hanno contribuito alla rivalutazione dei dati bioarcheologici soprattutto nell'ottica della condivisione, si pensi a titolo esemplificativo all'adozione dei *Linked Open Data* e agli effetti benefici sulla condivisione di dati e risultati [19: 2-3]. L'utilizzo di risorse informatiche in fase di documentazione ed elaborazione dei dati archeozoologici, non è tuttavia esente da problematiche di ordine metodologico e applicativo, insite nella necessità di adattamento dello strumento informatico allo stato della documentazione zooarcheologica, ed in particolare quando ci si rapporta alla documentazione pregressa, i cosiddetti *legacy data*, che spesso sono caratterizzati da una forte eterogeneità [17: 56].

3. CASI STUDIO

Si presentano, in questa sede, due casi studio inerenti le classi di materiali "minori" di cui si è detto, nei quali le problematiche documentali che li caratterizzano sono state affrontate da un punto di vista metodologico, proponendo, al contempo, delle soluzioni pratiche per il loro trattamento informatico.

Il primo riguarda la gestione dei dati relativi agli *small finds* provenienti dai due siti di Festòs e Haghia Triada a Creta, appartenenti cronologicamente ad una fase che va dal Neolitico Finale al Tardo Minoico IB (3650/3500-1425 a.C. ca.). Gli aspetti più problematici sono legati alla natura dei reperti ed alla lunga storia delle ricerche nei due siti, entrambi fattori che hanno contribuito ad una sostanziale mancanza di sistematicità nella raccolta dei dati e alla moltiplicazione degli stessi. Questo si evince, in particolar modo, nella variabilità delle attribuzioni tipologiche e funzionali assegnate a questa categoria di reperti, soprattutto qualora siano stati più volte oggetto di studio e verifica, secondo il principio della revisione degli studi proposto da La Rosa con il "riscavare lo scavato e rileggere il già letto" [10]. Per la gestione informatica è stato necessario effettuare una analisi critica e una comparazione sistematica dei *legacy data*, finalizzata alla messa in luce di tutte le incongruenze e ad evitare il rischio di perdere, sottostimare o tralasciare informazioni. Constatata la necessità di usare un approccio in grado di conservare l'incertezza e la variabilità del dato, si è scelto di tenere traccia non solo di tutte le fonti, dalle più vecchie alle più recenti, ma anche, per ciascuna fonte, di tutte le interpretazioni che sono state date nel tempo per ogni reperto. Il database relazionale realizzato gestisce, quindi, per il sito di Festòs e Haghia Triada rispettivamente 1221 e 626 *small finds*, 2090 e 1213 fonti, 2208 e 1314 diverse attribuzioni tipologiche, prendendo in

[2] Fino a questo punto, si attribuisce la responsabilità a Marianna Figuera; il resto del paragrafo è firmato da Erica Platania.

considerazione anche tutti i casi in cui la stessa fonte è incerta e fornisce una pluralità di attribuzioni. Grazie all'uso della logica *fuzzy*, metodo matematico capace di "sfocare" le rigide regole informatiche del vero o falso [8], si è, quindi, elaborato il concetto di probabilità di appartenenza di un reperto ad una tipologia specifica e si è calcolato un coefficiente di affidabilità di ciascun enunciato presente nelle fonti. Proprio la risoluzione dei problemi metodologici insiti nel caso studio ha condotto verso il trattamento dei concetti di rilevanza della fonte e del tipo di fonte e di attendibilità dell'attribuzione tipologica proponendo, quindi, un possibile approccio al riuso consapevole dei dati. I benefici nell'utilizzo di questo approccio sono immediatamente evidenti dal punto di vista archeologico, infatti è possibile distinguere i reperti che presentano un indice di affidabilità basso, suggerendo per essi una revisione dei precedenti studi; analizzare nuovamente i dati al fine di modificare l'interpretazione funzionale dei reperti e, di conseguenza, anche di interi contesti archeologici; tenere conto di tutte le fonti durante la fase di ripubblicazione e di rianalisi dei dati [5, 6: 103-106][3].

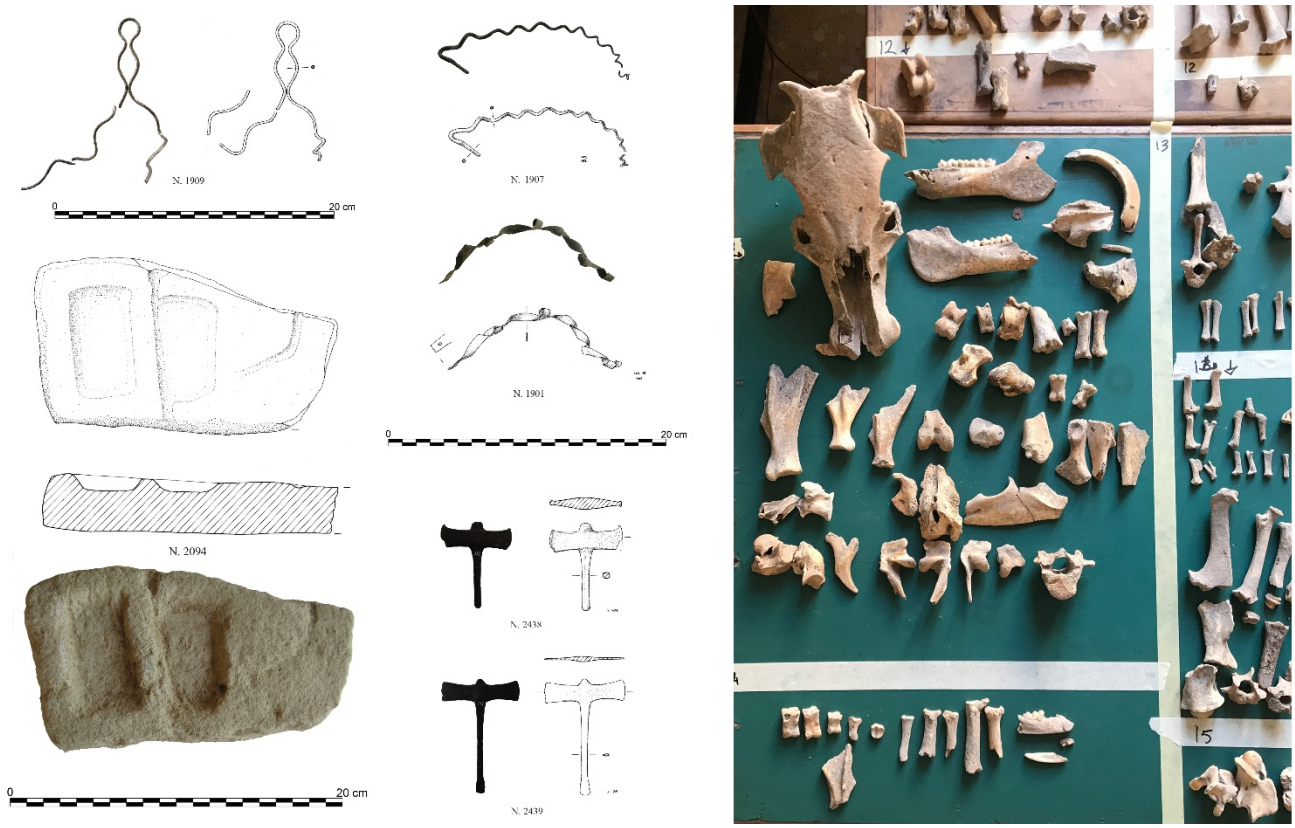


Figura 1. Alcuni esempi di reperti archeologici cd. "minori": a sinistra small finds *dal sito di Haghia Triada (ID_1909 e ID_1907 elementi filiformi; ID_2094 matrice litica; ID_1901 nastro in bronzo; ID_2438 e ID_2439 doppie asce miniaturistiche; foto di Marianna Figuera, disegni di Giuliano Merlatti), a destra resti osteologici animali in fase di studio, dal vano 26 del sito di Calaforno (RG), suddivisi per US di provenienza (foto di Erica Platania).*

Il secondo caso studio ha come oggetto la ricostruzione dei fenomeni pastorali in Sicilia orientale nella preistoria, nel periodo compreso tra il Neolitico e l'età del Bronzo antico (6200-1450 a.C. ca.) partendo dallo studio della documentazione archeozoologica edita, una fonte a lungo ritenuta secondaria negli studi di preistoria dell'isola, ma dotata di un alto potenziale informativo se opportunamente interrogata [17], integrata allo studio *ex novo* di resti faunistici provenienti da indagini di scavo recenti [18]. Durante il lavoro di ricognizione dei *legacy data* sono emerse alcune criticità che hanno determinato l'approccio metodologico utilizzato per la ricerca ed il ricorso a strumenti di archiviazione digitale per la normalizzazione e gestione dei dati. In primo luogo, tralasciando il problema legato alla scarsa consistenza in termini numerici di ricerche recenti, ciò che è subito emerso è la diffusa mancanza di strumenti di archiviazione digitale condivisibili, che deriva in parte dall'eterogeneità degli approcci metodologici utilizzati e che si riflette nella mancanza di standardizzazione dei dati, nella frammentazione delle informazioni e nella conseguente difficoltà di comparazione dei risultati e quindi di riuso degli stessi. In due casi è stato possibile disporre di dati archiviati su fogli di calcolo, che sono stati uniformati ai criteri utilizzati nell'elaborazione del *dataset* per i reperti inediti, per il resto dei casi sono stati estrapolati dalle pubblicazioni esclusivamente i dati quantitativi, rimandando ai singoli autori per l'interpretazione generale dei

[3] Fino a questo punto, si attribuisce la responsabilità a Marianna Figuera; il resto del paragrafo è firmato da Erica Platania.

campioni osteologici in riferimento a singoli contesti, essendo questa frutto di un'ampia gamma di informazioni non sempre disponibili nelle pubblicazioni. Al termine del lavoro di revisione si è potuto disporre di un campione costituito da 19 siti archeologici variamente distribuiti nell'area orientale dell'isola, in territorio etneo ed ibleo, afferenti alle moderne provincie di Catania, Siracusa e Ragusa. È stato quindi elaborato un database relazionale in grado di permettere la gestione dei dati con un buon livello di controllo. La progettazione del database ha dovuto, quindi, tenere in debita considerazione l'eterogeneità dei dati editi e la possibilità di integrazione di questi con i dati inediti che offrono inevitabilmente un maggior grado di dettaglio.

4. IL PROGETTO STORAGE

Il progetto interdipartimentale "Storage. Dai dati al Web" – cui hanno contribuito umanisti (DISUM) e informatici (DMI) dell'Università di Catania – dedicato al problema della raccolta, archiviazione, gestione e comunicazione digitale dei dati in ambito archeologico e storico-artistico, è stato il contesto ideale per la ripresa e messa a punto delle problematiche metodologiche affrontate separatamente nei due casi studio sopra descritti. Entrambi, infatti, presentavano una serie di aspetti comuni: 1) potenziale informativo dei reperti sottostimato; 2) gestione problematica dei cosiddetti *legacy data*; 3) mancanza di standardizzazione terminologica; 4) volontà di preservare tutte le fonti, evitando la perdita di informazioni; 5) necessità di creare un metodo di valutazione dell'affidabilità dei dati. Comune era, inoltre, la volontà di creare uno strumento che potesse uniformare quanto più possibile il quadro eterogeneo della documentazione disponibile, aperto a successive implementazioni, in grado di rispondere ad interrogazioni specifiche, coerentemente agli obiettivi precipui delle ricerche.

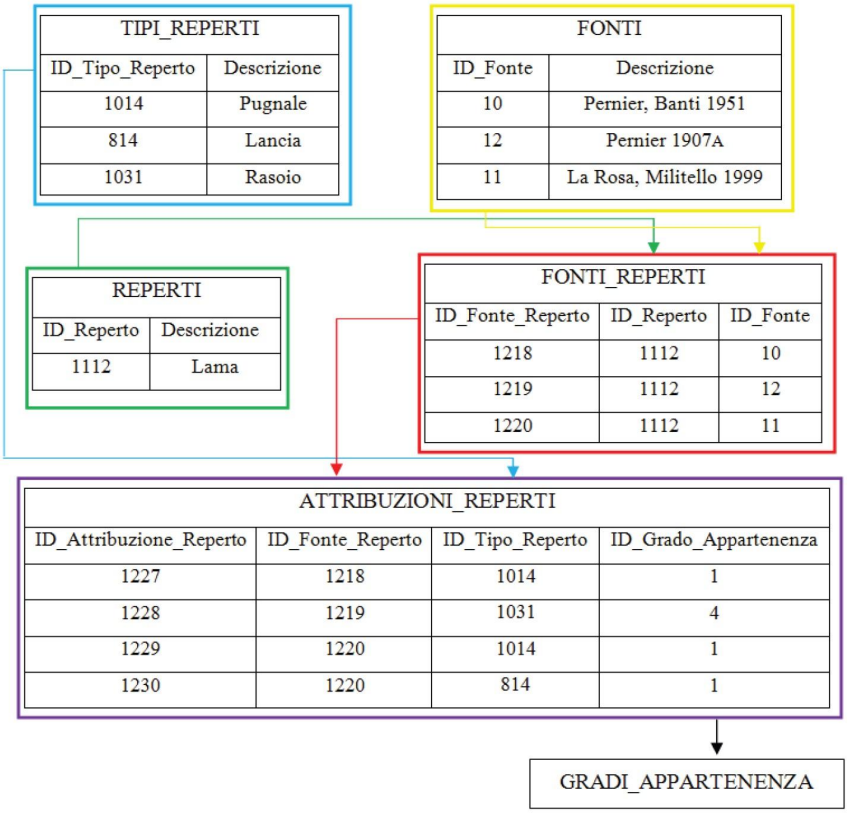
In particolare, per la gestione degli *small finds* il database relazionale realizzato ha un modello logico imperniato sulla tabella Reperti, mentre per le restanti 45 tabelle è possibile distinguere fra "entità di gestione", con tutte le informazioni di dettaglio, ed "entità dizionario", che contengono i glossari di riferimento. Le informazioni sono organizzate in tre macrocategorie focalizzate sui concetti di localizzazione (Siti, Aree_rinvenimento, Attività, Riferimenti_stratigrafici, ecc.), caratteristiche (Cronologie, Dimensioni, Materiali_reperti, Classi_reperti, Tracce_uso, Colori, Caratteristiche, ecc.) e documentazione (Soggetti, Riferimenti_bibliografici, Indici_reperti, Immagini_reperti, Fonti, Attribuzioni_reperti, Gradi_appartenenza, Tipi_reperti, Certificazioni_dati, ecc.). Di quest'ultimo gruppo fanno parte anche le 9 tabelle in cui è possibile gestire le informazioni legate alle fonti utilizzando, come già detto, l'approccio *fuzzy* che permette di conservare l'incertezza e la variabilità del dato, e di supportare l'affidabilità dell'attribuzione archeologica [5]. Nello specifico, l'archivio Gradi_appartenenza permette la definizione di diverse scale di valori (attendibilità dell'attribuzione tipologica, affidabilità del deposito archeologico, rilevanza della fonte, rilevanza del tipo fonte) espresse tramite etichette nel campo 'Descrizione', implementabili da ciascun ricercatore in base alle proprie esigenze metodologiche, alle quali corrispondono coefficienti numerici decimali (compresi tra 0 e 1) nel campo 'Peso'. Per la risoluzione dei conflitti generati dalla gestione di fonti multiple si è dovuto, infatti, valutare alcuni parametri: la recenziorità della fonte, l'autorevolezza scientifica delle affermazioni (desunta dai dati scientifici e dalle analisi di dettaglio presentate a supporto), il grado di approfondimento della fonte (con studi a supporto; di tipo compilativo; di tipo preliminare; notazione ragionata; notazione immediata; ecc.), il grado di certezza espresso dall'autore all'interno della stessa attribuzione e la coerenza con precedenti interpretazioni. Nel caso, ad esempio, del reperto ID_1112 sono state date nel tempo tre differenti attribuzioni tipologiche: 'Rasoio' nel 1907 [15: 281], 'Pugnale' nel 1951 [16: 181], 'Pugnale' o 'Lancia' nel 1999 [11: 262], il metodo *fuzzy* permette di tenere traccia di ciascuna di esse, di assegnare un valore di 'probabilità di appartenenza' ad ogni tipologia e, prendendo in considerazione tutti i criteri di affidabilità sopra detti, calcola un coefficiente di affidabilità per ogni enunciato, inteso come citazione testuale della descrizione del reperto all'interno di un qualsivoglia documento [6: 98-103] (vd. Fig. 2)[4].

Il database realizzato per la gestione dei dati faunistici della preistoria della Sicilia orientale, è articolato in due *dataset,* denominati "Siti" e "Reperti" destinati a ospitare dati di origine diversa. Il *dataset* "Siti" è stato progettato in conformità alla tipologia delle informazioni disponibili nella documentazione edita e riunisce le informazioni inerenti la geomorfologia (altitudine, ambiente, fascia altimetrica, idrologia); la posizione (provincia, latitudine e longitudine); la cronologia (datazioni radiometriche, datazione assoluta e relativa); la facies culturale; la tipologia del sito e del contesto; il numero totale di elementi osteologici animali rinvenuti; il grado di affidabilità dei dati; il totale degli elementi determinati e non determinati; il totale degli elementi suddivisi per specie di appartenenza e la relativa percentuale di rappresentazione nell'intero assemblaggio faunistico; note e bibliografia di riferimento. Il *dataset* "Reperti" ha rappresentato in fase di progettazione la sfida più impegnativa, questo infatti doveva avere una struttura in grado di adattarsi alla particolare categoria dei dati trattati (archeofaunistici), ed essere aperta a modifiche e ampliamenti essendo destinato all'archiviazione

[4] Fino a questo punto, si attribuisce la responsabilità del paragrafo a Marianna Figuera.

dei dati provenienti dagli studi inediti. L'archivio è stato pensato e costruito attorno al singolo reperto e i suoi attributi sono organizzati in due gruppi di informazioni:

- riferimenti contestuali (id_reperto, id_sito, saggio, area, quadrato, US, cronologia, data)
- analisi del reperto (tassonomiche, anatomiche, tafonomiche e morfometriche).



TIPI_REPERTI

ID_Tipo_Reperto	Descrizione
1014	Pugnale
814	Lancia
1031	Rasoio

FONTI

ID_Fonte	Descrizione
10	Pernier, Banti 1951
12	Pernier 1907A
11	La Rosa, Militello 1999

REPERTI

ID_Reperto	Descrizione
1112	Lama

FONTI_REPERTI

ID_Fonte_Reperto	ID_Reperto	ID_Fonte
1218	1112	10
1219	1112	12
1220	1112	11

ATTRIBUZIONI_REPERTI

ID_Attribuzione_Reperto	ID_Fonte_Reperto	ID_Tipo_Reperto	ID_Grado_Appartenenza
1227	1218	1014	1
1228	1219	1031	4
1229	1220	1014	1
1230	1220	814	1

GRADI_APPARTENENZA

Figura 2. Tabelle del database coinvolte nella gestione delle informazioni legate alle fonti utilizzando il metodo fuzzy *(elaborazione di Marianna Figuera.*

In linea con gli obiettivi di "Storage" si è dunque partiti dai due database sopra descritti al fine di unificare e uniformare in un unico sistema di gestione dei dati quanto prodotto nei due rispettivi ambiti di applicazione, testando inoltre le potenzialità del metodo *fuzzy* per la gestione dei dati archeozoologici sulla pastorizia nella Sicilia orientale. È emerso che anche in questo caso la *fuzzy* può essere utilizzata per gestire le cd. "etichette", ovvero categorie predefinite per cui è possibile calcolare la probabilità di appartenenza [8: 99-103]. Quindi trova una prima applicazione per la valutazione dell'affidabilità dei dati relativi alla interpretazione archeologica del contesto e quella generale del record archeozoologico, aspetti per cui sono disponibili più fonti, anche di natura diversa, sullo stesso sito. Un secondo campo particolarmente pertinente agli obiettivi della ricerca è quello relativo all'età di morte degli individui. Tale dato infatti è di particolare importanza per ricostruire le modalità di sfruttamento delle risorse animali [14] e al contempo uno dei più problematici da ottenere sulla base della documentazione esistente, in quanto spesso presenta un notevole grado di incertezza. Questa deriva in primo luogo dalla stessa natura dei resti osteologici, dalle condizioni di conservazione del record e dalle metodologie che è possibile utilizzare per la determinazione. I dati che è possibile ottenere dall'analisi della fusione delle epifisi articolari, così come dall'osservazione dello stato di eruzione e usura dei denti restituisce intervalli d'età, talvolta piuttosto ampi, tramite cui stabilire l'appartenenza a classi d'età prestabilite (giovane, adulto, senile, ecc.). Tuttavia, lo stato della documentazione spesso non permette di attribuire con sufficiente certezza l'appartenenza di un individuo/elemento osteologico ad una classe d'età, con il risultato che tali elementi vengono scartati in fase di elaborazione. L'applicazione della logica *fuzzy* permette in questo caso di attribuire un grado di appartenenza ad una o più classi d'età, con il vantaggio di utilizzare all'interno delle analisi elementi precedentemente esclusi, poter ampliare notevolmente la base di dati di partenza contribuendo quindi ad una migliore interpretazione del record archeozoologico [9][5].

Grazie ad una preliminare analisi dei requisiti si è giunti, quindi, alla modellazione dei due domini: questa fase ha comportato un ulteriore processo di normalizzazione terminologica, e soprattutto concettuale, che tenesse conto della

[5] Fino a questo punto, si attribuisce la responsabilità del paragrafo a Erica Platania.

tendenza, che entra in gioco nel trattamento informatico del dato, ad annullare, o quantomeno minimizzare, le differenze esistenti fra i *dataset*. È stato quindi creato un modello concettuale unitario che risponde ai requisiti di completezza, correttezza, leggibilità e minimalità (o non ridondanza). Per documentare le informazioni presenti nel sistema si è realizzato un *Entity Relationship diagram* che ne rappresenta la struttura attraverso le entità e le relazioni tra esse. L'entità fulcro è "Reperto", cui sono collegate, tramite relazioni basate sui principi di cardinalità e obbligatorietà, le altre entità suddivise in tre categorie funzionali di informazioni: 1) aspetti legati alla localizzazione e al contesto, 2) caratteristiche e cronologia del reperto, 3) dati relativi alla documentazione (vd. Fig. 3).

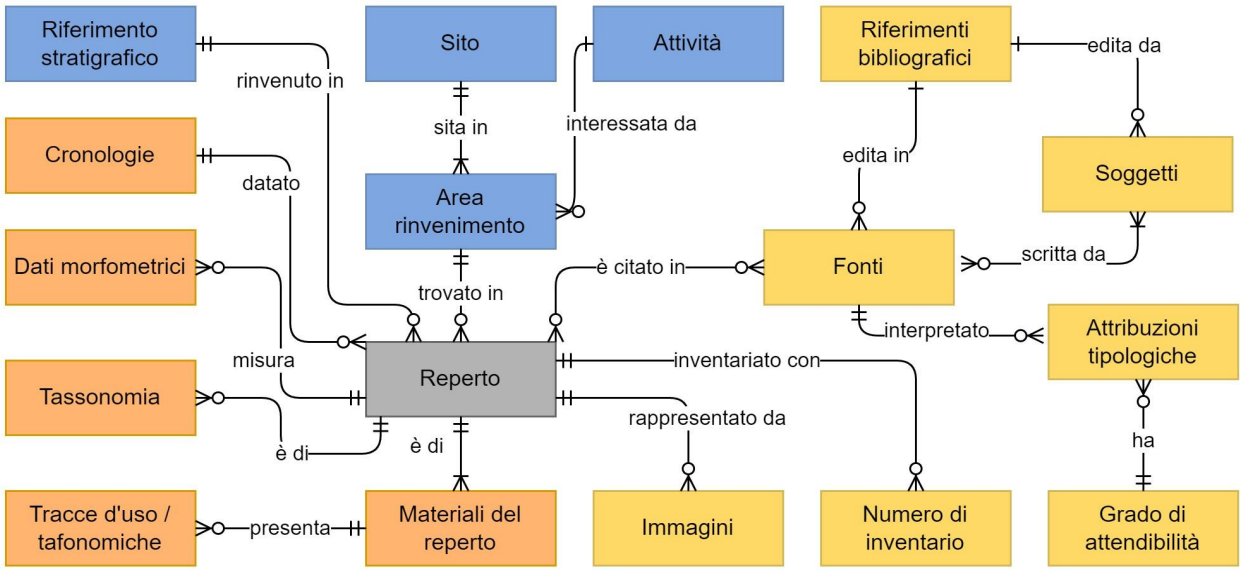


Figura 3. E-R diagram *del sistema realizzato nell'ambito del progetto "Storage" con l'entità fulcro "Reperto" (in grigio) e le entità riguardanti la localizzazione (in azzurro), le caratteristiche (in arancio) e la documentazione (in giallo) (elaborazione di Marianna Figuera e Erica Platania).*

Partendo dal modello concettuale si è quindi passati a strutturare il modello logico, integrando i *dataset* dei dati faunistici con quelli degli *small finds*, mantenendo la struttura relazionale. In parallelo, al lavoro di normalizzazione terminologica, già effettuato in seno ai due progetti, si è affiancato l'adeguamento del sistema alle normative dettate dall'ICCD, prendendo atto della necessità di standardizzazione, altro elemento chiave di "Storage". Obiettivo ultimo è stata la migrazione delle strutture e dei dati in un database MySQL, che sostituisce quelli precedenti (Oracle 10 XE in versione freeware per fini di ricerca e Microsoft Access), utilizzando una infrastruttura controllata e con backup garantiti e preservando le funzionalità presenti quali le procedure di calcolo degli indici *fuzzy*.
Per rispondere ai requisiti di accessibilità dei dati è attualmente in fase di elaborazione l'interfaccia web atta a garantire il monitoraggio dello *status quo* della ricerca e a fornire alla comunità scientifica la possibilità di consultazione e verifica dei dati.

5. CONCLUSIONI

Il progetto "Storage" ha permesso, grazie alla collaborazione fra specialisti di ambito umanistico e informatico, di affrontare alcuni dei nodi metodologici che caratterizzano il rapporto tra le due discipline. In particolare, i due casi studio, relativi alla gestione informatica dei dati sugli *small finds* da Festòs e Haghia Triada a Creta e sui fenomeni pastorali della Sicilia orientale in età preistorica, hanno offerto la possibilità di indagare aspetti e problematiche specifiche legate alla natura di queste classi di materiali archeologici, erroneamente considerate secondarie nell'ambito della ricerca archeologica. Il loro potenziale informativo, al contrario, è stato valorizzato proprio grazie ad una gestione informatica che tiene conto di aspetti quali: la riconsiderazione di tutti i *legacy data*; la sistematizzazione e normalizzazione della struttura dei dati e dei vocabolari di riferimento; la flessibilità per il trattamento di informazioni di natura diversa, da contesti di differente ambito cronologico e geografico; la possibilità di preservare l'incertezza del dato grazie all'applicazione della metodologia *fuzzy*; la creazione di un sistema di validazione dell'affidabilità delle fonti.
Il sistema unitario sviluppato risponde ad alcuni requisiti fondamentali quali l'accessibilità – attraverso lo sviluppo dell'interfaccia web che permette la consultazione e la verifica dei dati immessi da parte della comunità scientifica e non –

e il riuso consapevole delle fonti preservando l'origine dell'informazione in modo da evitare semplificazioni e da lasciare traccia di tutti i dati, anche contrastanti, permettendo al sistema di divenire una sorta di "contenitore di memorie".

6. RINGRAZIAMENTI

Parte della ricerca si è svolta nell'ambito delle tesi di ricerca dottorale in Studi sul Patrimonio Culturale, XXIX e XXXII ciclo, dell'Università di Catania. Il lavoro è finanziato per gli anni 2020-2022 dal progetto interdipartimentale (DISUM-DMI) "Storage. Dai dati al Web" programma Pia.Ce.Ri dell'Università di Catania (Marianna Figuera, Erica Platania) e dal 2023 nell'ambito del progetto CHANGES, Spoke 6, History, Conservation, Restoration of Cultural Heritage (Erica Platania).

BIBLIOGRAFIA

[1] Allison, Penelope. «Dealing with Legacy Data - An introduction». *Internet Archaeology* 24 (2008). https://doi.org/10.11141/ia.24.8.

[2] Burström, Myrberg. «Things in the Eye of the Beholder: A Humanistic Perspective on Archaeological Object Biographies». *Norwegian Archaeological Review* 47, fasc. 1 (2014): 65–82.

[3] Caraher, William. «Slow archaeology: technology, efficiency and archaeological work». In *Mobilizing the Past for a Digital Future: The Potential of Digital Archaeology*, a cura di E. W. Averett, J. M. Gordon, e D. B. Counts, 421–41. North Dakota: The Digital Press at the University of North Dakota, 2016.

[4] De Grossi Mazzorin, Jacopo. *Archeozoologia, lo studio dei resti animali in archeologia*. Editori Laterza. Roma-Bari, 2008.

[5] Figuera, Marianna. «Database management e dati archeologici: standardizzazione e applicazione della Logica Fuzzy alla gestione delle fonti e delle attribuzioni tipologiche.» *Archeologia e Calcolatori* 29 (2018): 143–60.

[6] Figuera, Marianna. «Un sistema per la gestione dell'affidabilità e dell'interpretazione dei dati archeologici. Percezione e potenzialità degli small finds: il caso studio di Festòs e Haghia Triada». *Praehistorica Mediterranea* 8 (2020).

[7] Gnesi Bartolani, Diego. «La catalogazione dei Beni Archeologici», dispense di Informatica applicata all'Archeologia, 2012.

[8] Hermon, Sorin, e Franco Niccolucci. «La logica fuzzy e le sue applicazioni alla ricerca archeologica». *Archeologia e Calcolatori* 14 (2003): 97–110.

[9] Hernon, Sorin, Franco Niccolucci, Francesca Alhaique, Maria Rosa Iovino, e Valentina Leonini. «Archaeological typologies – an archaeological fuzzy reality.» In *Enter the Past, the E – way into the Four Dimensions of Cultural Heritage*, 30–34. B.A.R. International Series. (a cura di) Wien Magistrat der Stadt. Oxford: Archaeopress, 2004.

[10] La Rosa, Vincenzo. «Ξανασκάβοντας το σκαμμένο˙ επιστημονική συνείδηση ή ασυνείδησία; Η εμπειρία της Αγίας Τριάδας (Κρήτη)». In *The Prehistoric Research in Greece and Its Perspectives. Theoretical and Methodological Considerations*, (a cura di) E. Vulgari, 165–69. Θεσσαλονίκη: University Studio Press, 2003.

[11] La Rosa, Vincenzo, e Pietro Militello. «Caccia, guerra o rituale? Alcune considerazioni sulle armi minoiche da Festòs e Haghia Triada». In *POLEMOS: Le contexte guerrier en Égée á l'âge du Bronze*, (a cura di) Robin Laffineur, 19:241–64. Liège-Austin: Aegaeum, 1999.

[12] Lucas, Gavin. *Critical Approaches to Fieldwork. Contemporary and historical archaeological practice*. London-New York: Routledge, 2001.

[13] Niccolucci, Franco, e Sorin Hermon. «Expressing reliability with CIDOC CRM». *International Journal on Digital Libraries* 18 (2017): 281–87.

[14] Payne, Sebastian. «Kill-off patterns in sheep and goats: the mandibles from As ̧van Kale». *Anatolian Studies* 33 (1973): 65–81.

[15] Pernier, Luigi. «Lavori eseguiti dalla Missione Archeologica Italiana a Creta (2 aprile - 12 settembre 1906)». *Rendiconti dell'Accademia dei Lincei* XVI (1907): 257–303.

[16] Pernier, Luigi, e Luisa Banti. *Il palazzo minoico di Festòs, Il secondo palazzo*. Vol. II. Roma: Istituto Poligrafico dello Stato, 1951.

[17] Platania, Erica. *Fonti archeozoologiche e strategie di sussistenza nella Preistoria della Sicilia orientale*. Vol. 7. Syndesmoi. Catania-Varsavia: Università degli studi di Catania, 2021.

[18] Platania, Erica. «Gli esordi della pastorizia nella Sicilia preistorica - il contributo dell'archeozoologia alla comprensione dei mutamenti nel sistema di sfruttamento delle risorse animali dal neolitico all'età del bronzo nella Sicilia sud-orientale». In *La Sicilia Preistorica. Dinamiche interne e relazioni esterne*, a cura di Pietro Militello, Fabrizio Nicoletti, e Rosalba Panvini, 187–96. Palermo: Regione siciliana, Assessorato dei beni culturali e dell'identità siciliana, Dipartimento dei beni culturali e dell'identità siciliana, 2021.

[19] Spyrou, Anna, Gary Nobles, Angelos Hadjikoumis, Allowen Evin, Arden Hulme-Beaman, Canan Çakirlar, Carly Ameen, et al. «Digital Zooarchaeology: State of the art, challenges, prospects and synergies». *Journal of Archaeological Science: Reports* 45 (2022): 1–9. https://doi.org/10.1016/j.jasrep.2022.103588.

[20] Steel, Louise. *Materiality and Consumption in the Bronze Age Mediterranean*. Vol. 7. Routledge Studies in Archaeology. New York: Routledge, 2013.

Gestire l'immateriale. ConTesti sensoriali a servizio del Patrimonio archeologico

Serena D'Amico
Dipartimento di Scienze Umanistiche, Università di Catania, Italia - seredamico@phd.unict.it

ABSTRACT

La ricerca in oggetto prende le mosse dalle più recenti problematiche emerse nell'ambito dell'Archeologia cognitiva ed in particolare dell'Archeologia sensoriale e propone un metodo di analisi degli spazi, ancora in fase sperimentale, che adotta due principali strumenti di misurazione: il corpo umano e i software. Obiettivo specifico è quello di comprendere i contesti archeologici attraverso l'esperienza sensoriale di un gruppo selezionato di persone, nel tentativo di individuare delle analogie comportamentali legate all'uso degli spazi. Mutuando dalle scienze cognitive metodi e tecniche di acquisizione dei dati qualitativi, questi verranno poi trascritti e analizzati con l'ausilio del software ATLAS.ti© consentendo di approfondire i contenuti dell'esperienza e d'individuare le dimensioni di senso più profonde. Sebbene ATLAS.ti© risulti ampiamente utilizzato in ambito sociologico, esso è pressoché sconosciuto nel campo della ricerca archeologica. I risultati attesi prevedono una lettura dei contesti archeologici più ampia, che nella ricostruzione della complessa realtà culturale del passato tenga conto dei dati immateriali sottesi agli spazi come ulteriori fonti informative.

Le finalità sono molteplici: stimolare innanzitutto la ricerca e lo sviluppo di strumenti e tecnologie sempre più adatti all'applicazione archeologica; rivelare le reali potenzialità di uno strumento di analisi quali-quantitativa praticamente mai utilizzato in archeologia; rendere sempre più partecipe l'Italia del dibattito scientifico europeo in materia alla luce delle sfide future che la attendono.

PAROLE CHIAVE

Sensory archaeology; patrimonio immateriale; Digital Humanities; ATLAS.ti; analisi testuali.

1. INTRODUZIONE: ARCHEOLOGIA COGNITIVA E SENSORIALE

La Cognitive Archaeology costituisce una delle branche più contemporanee dell'archeologia, ciononostante possiede solide basi teoriche, metodologiche, ontologiche ed epistemologiche sviluppate in un cinquantennio di ricerche ed una forza scientifica che risiede nell'aver mutuando teorie, metodi e modelli interpretativi da altri ambiti disciplinari (neuroscienze, psicologia, filosofia, antropologia). Nata dal fertile dibattito fra archeologi processualisti e post-processualisti, i presupposti su cui si fonda riconoscono un legame profondo che intercorre fra cognizione e materialità: dalla creazione di oggetti all'utilizzo del linguaggio, dalla produzione artistica alla manipolazione degli spazi, in sostanza qualsiasi traccia materiale ed intellettuale di un popolo, la cultura nel suo complesso, tutto risulta essere il risultato di processi cognitivi di acquisizione e comprensione di informazioni attraverso il pensiero, i sensi, l'esperienza.

Dei due principali indirizzi di ricerca della *Cognitive Archaeology* l'ECA (*Evolutionary cognitive archaeology*) [1, 10] e l'ICA (*Ideational cognitive archaeology*), è in quest'ultimo che fra gli anni '90 e i primi anni 2000 in Europa e America ha preso campo la *Sensory archaeology* [12]. Focalizzata sulla percezione come processo essenziale di comprensione del mondo, indaga l'esperienza come somma sinestetica di molteplici stimoli percettivi per raggiungere la comprensione del sensorium delle culture del passato. L'avvio di tale disciplina[1], inizialmente ostacolato dal predominio dell'approccio processuale, ha creato un'aura di grande e generale scetticismo, specie riguardo alla fenomenologia sostenuta da Christopher Tilley [13]; scetticismo in parte superato grazie al lavoro svolto dal Sensoria Research Team della Concordia University di Montreal (CONSERT)[2] diretto da David Howes.

Da allora in poi i sensi sono diventati oggetto di studio e mezzo di indagine, consentendo di rovesciare modelli interpretativi fino ad allora mai messi in discussione (la gerarchia aristotelica dei cinque sensi, la supremazia della mente nel dualismo cartesiano 'mente-corpo', ecc.) e di formulare nuovi costrutti teorici (l'*embodiement*, la 4E *Cognition*, ecc.). Se da un lato la mole di volumi e articoli in rivista pubblicati nell'ultimo trentennio ha evidenziato la trasversalità dell'archeologia sensoriale e l'eterogeneità degli aspetti materiali coinvolti nella percezione (manufatti, strutture, paesaggi, ecc.), dall'altro ha però dimostrato l'inadeguatezza delle metodologie tradizionali nell'accedere alla dimensione sensoriale del passato.

[1] Orientata in senso fenomenologico, essa ha di fatto camminato in parallelo all'Antropologia dell'esperienza proposta da Victor Turner nel 1986 [7: 150-152].
[2] https://www.david-howes.com/senses/index.htm

Sebbene l'Italia non sia estranea alle riflessioni sul Patrimonio immateriale, il cui concetto normativo è ben consolidato e la moltitudine delle sue espressioni chiaramente indagabile [2], degli aspetti sensoriali si parla poco e solo da anni recenti. Quando è il Patrimonio archeologico ad essere chiamato in causa, il dialogo sul tema della percezione sensoriale si assesta sulla visualizzazione dello spazio antico attraverso esperienze sempre più immersive ed inclusive, ma che sembrano più mirate alla comprensione dell'uomo di oggi che non a quello del passato. Un dialogo a monte, genuinamente archeologico, incentrato cioè sulla definizione di metodologie e strumenti d'indagine del 'sentire' antico è in buona sostanza ancora debole, se non addirittura assente; lo conferma il fatto che in Italia un'Archeologia sensoriale non esiste. Eppure, l'ampia letteratura prodotta nel resto dei Paesi, specie quelli anglofoni, ha già rivelato le potenzialità dell'integrare la dimensione sensoriale nella più ampia interpretazione degli spazi e dei materiali archeologici.
Con queste premesse, i recenti sviluppi compiuti nel campo del digitale possono non soltanto stimolare il dibattito sul tema, ma anche denunciare la necessità di un'Archeologia sensoriale solida nei contenuti e nel metodo che rimanga al passo col progresso tecnologico, affinché essa stessa possa rendersi motrice di nuovi sviluppi informatici. In questa direzione si muove il presente lavoro, ancora nella sua fase sperimentale, sviluppato in una più ampia attività di ricerca dottorale che indaga l'*agency* esercitata dagli spazi in grotta ed il ruolo giocato dalla cognizione nel loro uso cultuale e funerario durante la Preistoria[3].

2. SPAZI DI SENSO

L'archeologia sensoriale sintetizza molte delle sfide dell'archeologia contemporanea affrontando tematiche e problematiche di frontiera difficili da approcciare senza multidisciplinarietà; ad aver beneficiato delle sperimentazioni di archeologia sensoriale sono stati soprattutto i paesaggi e i monumenti, il cui studio è stato finora affrontato attraverso due principali approcci: fenomenologico e digitale.
Il primo si avvicina allo spazio del passato mediante analogie dell'esperienza, valutata con soggettività per come appare oggi. Gli strumenti di indagine sono diversi e combinabili (metodo riflessivo, inventariale, sperimentale, descrizione densa, scrittura creativa, questionari, ecc.). I suoi punti più deboli sono rappresentati dall'inferenza empirica e dalla mancanza di un metodo di riferimento che inquadri la realtà percepita all'interno di una definizione specifica o una categoria diagnostica, ossia entro parametri oggettivi; a ciò si aggiungono le inevitabili modifiche, avvenute nel corso dei secoli, dello stato originario degli spazi osservati. Nonostante ciò, l'approccio fenomenologico è dagli anni '90 ad oggi ampiamente utilizzato dagli archeologi [5, 11], in quanto continua a restituire le più vicine dimensioni di senso provate nella realtà fisica.
Analisi degli spazi affrontati in maniera più strutturata, vale a dire attraverso metodi quantitativi e statistici ma non ancora computazionali, sono quelli tentati negli anni '80, che rientrano da un lato nei cosiddetti "studi di visibilità non-GIS", attraverso cui sono stati valutati l'impatto visivo, le variabili di movimento, l'accessibilità, ecc., dall'altro nei metodi della Space Syntax, atta a indagare le relazioni tra ambiente costruito e comportamento umano attraverso modelli matematici predittivi di varia complessità [6]. Gli studi degli anni '80 hanno di fatto gettato le basi per il processo di digitalizzazione degli spazi archeologici che, parallelamente ai progressi compiuti nel campo della grafica computerizzata (software CAD, GIS, fotogrammetria, modellazione 3D, Virtual Reality, ecc.), si è focalizzata sul rendering di paesaggi e monumenti, restituiti non solo in termini visuali, ma anche acustici. Sebbene tale approccio sia più oggettivo, dunque più affidabile scientificamente, esso non è esente da problematiche, soprattutto di carattere tecnico.
L'approccio fenomenologico e quello digitale possiedono comunque un punto debole di fondo legato alla natura intrinseca degli spazi archeologi, specie dei paesaggi pluristratificati, dato che questi non restituiscono un momento specifico della storia bensì una realtà artificiale dove ciò che resta è mutilato dalle attività pre- e post-deposizionali [4].

3. PER UNA TRADUZIONE DIGITALE DELL'ESPERIENZA SENSORIALE

Il metodo adottato nella ricerca in oggetto tiene conto dei punti di forza e delle debolezze proprie di ogni approccio collocandosi in una posizione intermedia fra quello fenomenologico e quello digitale; ci si avvale cioè della combinazione di dati qualitativi e quantitativi.
Per ragioni di spazio non verrà qui trattata la fase dell'esperienza sensoriale sul campo. In linea generale essa prevede il coinvolgimento di un gruppo selezionato di persone sottoposte allo svolgimento di alcune attività all'interno di spazi archeologici ritenuti ottimali per la sperimentazione sensoriale. Ci si riferisce nello specifico ai contesti sotterranei (grotte naturali e artificiali, catacombe, ipogei, ecc.), che per le intrinseche qualità dei loro ambienti (carenza/assenza di luce, alto grado di umidità, ristrettezza degli spazi, pseudo-/irregolarità delle superfici, ecc.) si prestano maggiormente alla stimolazione dei recettori sensoriali (chemiocettori, meccanocettori, barocettori, termocettori, propriocettori, ecc.).

[3] Il lavoro rientra inoltre nell'ambito del progetto interdipartimentale (DISUM-DMI) dell'Università di Catania "Storage. Dai dati al web" (Programma Pia.Ce.Ri).

Sebbene questi producano segnali nervosi anche nei restanti contesti in superficie, la compromissione dei fotorecettori (responsabili della vista) comporta infatti una differente decodifica degli spazi da parte degli organi di senso[4]. Le medesime qualità possedute da tali ambienti, inoltre, se combinate in fase interpretativa al ruolo funzionale da essi svolto (funerario e cultuale) suggeriscono il potere evocativo e le suggestioni emotive che pur devono aver stimolato l'apparato sensoriale durante lo svolgimento di rituali e nella costruzione della memoria.

L'acquisizione dei dati qualitativi derivanti dall'esperienza avviene perlopiù sul campo e in minima parte da remoto attraverso strategie e metodi propri delle scienze cognitive (questionari non e semi-strutturati, interviste, focus group, descrizioni dense, scrittura creativa, ecc.). I dati raccolti, organizzati in cartelle tante quante sono le persone sottoposte all'esperienza sensoriale, presentandosi nella sostanza come dati grezzi (*raw data*) non possono tuttavia essere direttamente analizzati e trasformati in risultati di ricerca rigorosi.

È a questo punto della ricerca che subentra l'utilizzo di software per la trascrizione testuale e l'analisi dei dati. Rispondendo alla necessità di oggettivare l'esperienza sensoriale, l'uso di software di analisi quali-quantitativa si rivela estremamente utile alla ricerca in un'ottica sia di orientamento sia di convalida. Nell'ottica di orientamento gli strumenti di analisi possono indicare piste di ricerca talvolta inattese facendo emergere dimensioni semantiche e tematiche non sempre riconoscibili. Nell'ottica di convalida è invece possibile dimostrare con dati empirici ciò che si suppone, senza il rischio di sovrastimare i fenomeni esistenti, offrendo anche il vantaggio della sintesi grafica dei risultati raggiunti [3: 299-300]. Il mercato digitale offre oggi una vasta gamma di software di analisi quali-quantitativa che combinano strumenti statistici, tecnologie informatiche e risorse linguistiche, genericamente noti nel mondo anglofono con l'acronimo CAQDAS (*Computer Assisted Qualitative Data Analysis Software*).

Sulla base degli obiettivi della ricerca e di altri fattori non meno discriminanti (costi di abbonamento, tipologie e durata delle licenze, sistema operativo supportato, caratteristiche funzionali, qualità e facilità d'uso delle procedure statistiche, output grafici offerti, flessibilità di importazione dei dati, velocità di elaborazione e di esportazione dei risultati, grado di soddisfazione degli utenti, ecc.) si è scelto di adottare il software di analisi testuale ATLAS.ti©[5], sviluppato nel 1993 dal tedesco Thomas Muhr ed oggi utilizzato in diverse parti del mondo da importanti aziende ed organizzazioni pubbliche e private (Microsoft, Google, the United Nations, Lego, ecc.), oltreché da prestigiose università (Freie Universität Berlin, Berkeley University of California, Harvard Business School, ecc.) ed istituti di ricerca (American Institutes for Research, Pacific Research Institute, Spanish National Research Council, ecc.).

L'utilizzo di ATLAS.ti© è stato pensato dallo stesso Muhr come uno strumento di grande utilità anche per gli etnografi[6]. Nel campo della ricerca archeologica non è invece nota in letteratura nessuna applicazione. Eppure, nel 2003, ossia a dieci anni di distanza dall'uscita sul mercato del software, l'archeologo Steve Townend della London's Global University ne aveva già riconosciuto le potenzialità raccomandandolo alla comunità scientifica degli etnoarcheologi, ma anche a coloro che sono impegnati in ricerche di archeologia pubblica, in studi museali e più genericamente a chi lavora con il patrimonio culturale [14: 168-169][7]. Il software è noto anche ai sistemi bibliotecari del Massachussets Institute of Technology[8] e della New York University[9], che alla sezione online "Anthropology & archaeology" ne raccomandano l'uso assieme a pochi altri software QDA selezionati (NVivo, Taguette, Dedoose, QDA Miner, MaxQDA, QCoder) offrendo inoltre un'utile comparazione fra questi.

ATLAS.ti© risulta invece particolarmente utilizzato nelle ricerche di tipo psicosociale, appartenendo alla categoria dei *Theory Building Software*[10]; sia la terminologia adottata sia i principi del suo funzionamento sono sviluppati nell'ambito del "paradigma interpretativo" basandosi esplicitamente sulla *Grounded Theory*. Il metodo che la contraddistingue evita che il ricercatore sviluppi categorie di analisi su ipotesi precostituite (metodo induttivo); sarà infatti l'analisi stessa a far emergere le categorie di dati (metodo deduttivo), raccolti, codificati e analizzati progressivamente in una comparazione

[4] È stimato, infatti, che la visione coinvolga i cinque sensi per ca. l'80%, non rimanendo cioè confinata alla sola capacità di vedere, ma trascinando con sé un'ampia gamma di abilità percettive (ad es.: identificazione degli oggetti, comprensione uditiva e verbale, equilibrio, orientamento, ecc.) attraverso le quali l'individuo si rapporta con oggetti e spazi.

[5] ATLAS.ti© Scientific Software Development GmbH, Berlin: https://atlasti.com/.

[6] https://atlasti.com/trainings/navigating-ethnographic-research-with-atlas-ti-tools-and-techniques#availabilities

[7] "In these areas of the discipline the need to conduct effective and coherent qualitative research is fundamental, yet most archaeologists are not familiar with, or in many cases even aware, that there is software available to support such research. With more widespread use, ATLAS.ti could become as indispensable to the qualitative researcher as the Statistical Package for the Social Sciences (SPSS) – to which, incidentally ATLAS.ti exports – is to the quantitative researcher. The archaeological community would do well to explore its potentialities further." [14: 169].

[8] https://libguides.mit.edu/anthro/qda

[9] https://guides.nyu.edu/QDA/atlasti

[10] ATLAS.ti GmbH. *User Manual 23 Mac*. Berlin, 2023.

ciclica e reiterata [8: 284]. Se da un lato l'approccio ai dati testuali è sistematico, servendosi di un modello a rete, dall'altro l'impostazione del software conferisce flessibilità all'intero processo di analisi.
Il lavoro del software si articola secondo una procedura di Descrizione-Analisi-Interpretazione che prevede vari step:

1. La costruzione dell'Unità Ermeneutica (Hermeneutic Unit - HU): rappresenta l'insieme dei 'Documenti primari' (Primary Documents – PDs) (vd. Fig. 1); in essa confluisce tutta la documentazione derivante dall'esperienza sensoriale, nel caso specifico i file audio (nei formati MP4 e OGG) contenenti le descrizioni dense, dei quali il software effettua la trascrizione, e i file di testo (in formato DOCX) relativi ai questionari[11];

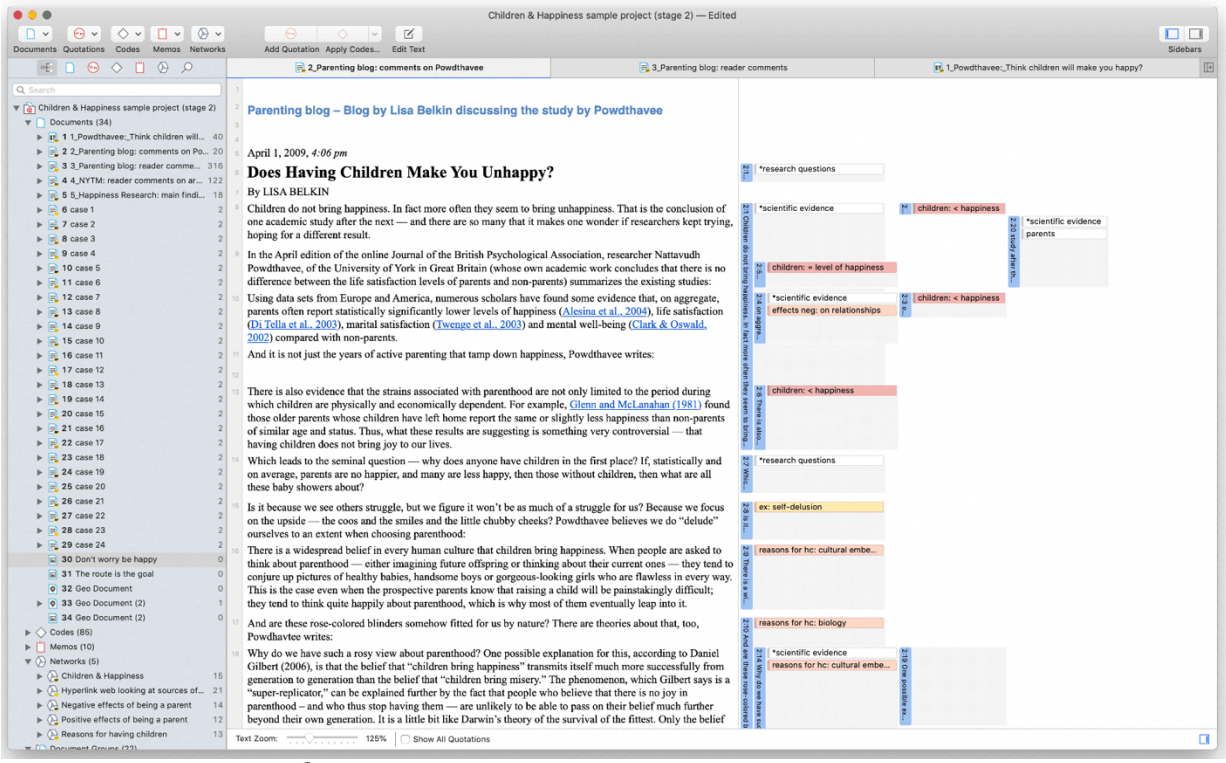


*Figura 1. Finestra di ATLAS.ti© relativa alla gestione dei documenti. Nell'esempio qui riportato la colonna di sn. (*navigator*) consente di visualizzare l'elenco dei documenti di lavoro, oltre ad altre* entities*. Sui singoli documenti di testo, visualizzabili nell'area centrale, è poi possibile creare i codici, aggiungere note, commenti e organizzarli per gruppi in base a delle categorie tematiche o semantiche create dal ricercatore*[12].

2. La codifica dei Documenti Primari: per rendere i dati più gestibili si scartano quelli che in nessun modo contribuiscono agli obiettivi della ricerca, si estraggono segmenti significativi e si effettuano soprattutto delle classificazioni/categorizzazioni dei dati in temi e pattern (detti "codici") che sintetizzano domini semantici, ossia specifici argomenti o temi (vd. Fig. 2). Dall'analisi dei testi relativi all'esperienza sensoriale negli spazi archeologici sotterranei si stanno ad es. codificando parole chiave come "paura", "angoscia", "pace", "conforto", "eco", "sacralità", ecc. all'interno di categorie come "dentro/fuori", "sensazioni positive/negative", "naturale/artificiale", "visibile/invisibile", "memoria", ecc. In sostanza, i *raw data* vengono trasformati in unità analizzabili costituendo la base da cui ricavare intuizioni e conclusioni.
   La codifica è un processo iterativo, per cui l'analisi dei dati e l'affinamento dei codici viene compiuto man mano che si procede. Il processo può essere effettuato manualmente o in maniera automatizzata a seconda delle necessità del ricercatore e avviene mediante la progressiva "codifica aperta", "codifica assiale" e "codifica selettiva". I codici così creati possono essere commentati (*edit comment*), organizzati in famiglie (*edit families*) e/o messi in rapporto fra loro (*link code to*); alla fine di tale processo, attraverso la funzione Query tool, è possibile interrogare i dati e rilevare la presenza di relazioni fra codici o code families nell'intera Unità Ermeneutica.

[11] Il software consente il caricamento di una vasta tipologia di documenti: di testo (.doc; .docx, .odt, .htm, .html, .txt, .ooxml); PDF generati dalla conversione, operata dal sistema, di Ebooks (.mobi), Excel (.xls, xlsx) e Libre Office Calc. (.ods), presentazioni Powerpoint (.ppt e .pptx) e Libre Office Impress (.odp), Visio (.vsd e .vsdx) e Libre Office Draw (.odg); immagini (.bmp, .gif, .jpeg, jpg, .png, t.if, .tiff); audio (.ac, .m4a, .mp3, .mp4); video (.avi, .m4v, .mov, .mp4); dati georeferenziati e commenti di social network (Facebook, Twitter, Instagram, YouTube, TikTok, VK, Twitch, Discord).
[12] https://doc.atlasti.com/ManualMac/Documents/DocumentsManaging.html (vd. nota 10).

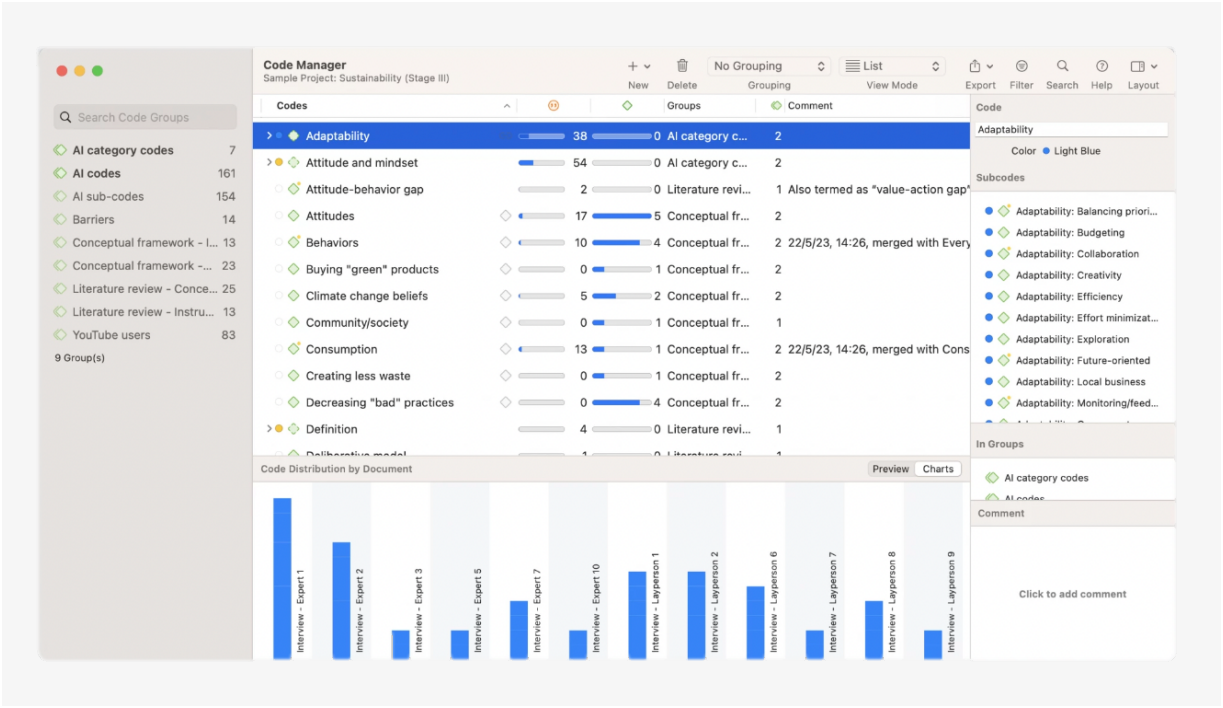


*Figura 2. Finestra di ATLAS.ti© relativa alla gestione dei codici. Nell'esempio qui riportato la colonna di sn. (*navigator*) consente di visualizzare l'elenco dei codici creati, differenziati per simbolo in base alla loro tipologia organizzativa (*Independent Code/Category Code/Subcode/Folder*). È poi possibile gestire singolarmente i codici, visualizzabili nell'area centrale, operare selezioni tramite filtri (a ds.) e avere una restituzione grafica della frequenza dei codici presenti nel gruppo di documenti (area centrale in basso)*[13].

3. La sintesi dei risultati: al termine del processo di codifica i risultati possono essere sintetizzati nel loro complesso o parzialmente in output testuali, visualizzazioni a finestra e rappresentazioni grafiche (vd. Fig. 3); in quest'ultimo caso il trattamento dei dati è di tipo quantitativo generando una Tabella dei Codici per i Documenti Primari che può essere salvata ed importata in eventuali altri software SPSS di analisi statistica.

4. CONCLUSIONI

La ricerca in oggetto, in corso di sperimentazione e giunta allo stadio di codifica dei Documenti Primari, sta innanzitutto confermando tutta l'efficacia degli strumenti di acquisizione sul campo dei dati sensoriali e l'enorme potenziale informativo in essi sotteso. L'utilizzo del software ATLAS.ti© sta rivelandosi uno strumento di enorme utilità ed efficacia per l'approfondimento e l'analisi dei contenuti semantici che consente non soltanto di far emergere la natura "incorporata e biologica" [9: 41] della percezione sensoriale per come si manifesta negli spazi archeologici, ma anche di configurare questi ultimi come reali dimensioni di senso.

In quest'ottica, inoltre, ATLAS.ti© sta dimostrando l'estrema versatilità dei suoi campi applicativi contribuendo così ad implementare una branca dello studio del passato finora lasciata in secondo piano, quella appunto dell'Archeologia sensoriale, che attesta la concreta possibilità di costruzione di ponti tra due realtà apparentemente inconciliabili: quella dell'esperienza soggettiva e quella della registrazione oggettiva del dato.

[13] https://doc.atlasti.com/ManualMac/Codes/CodingDataBasicConcepts.html (vd. nota 10).

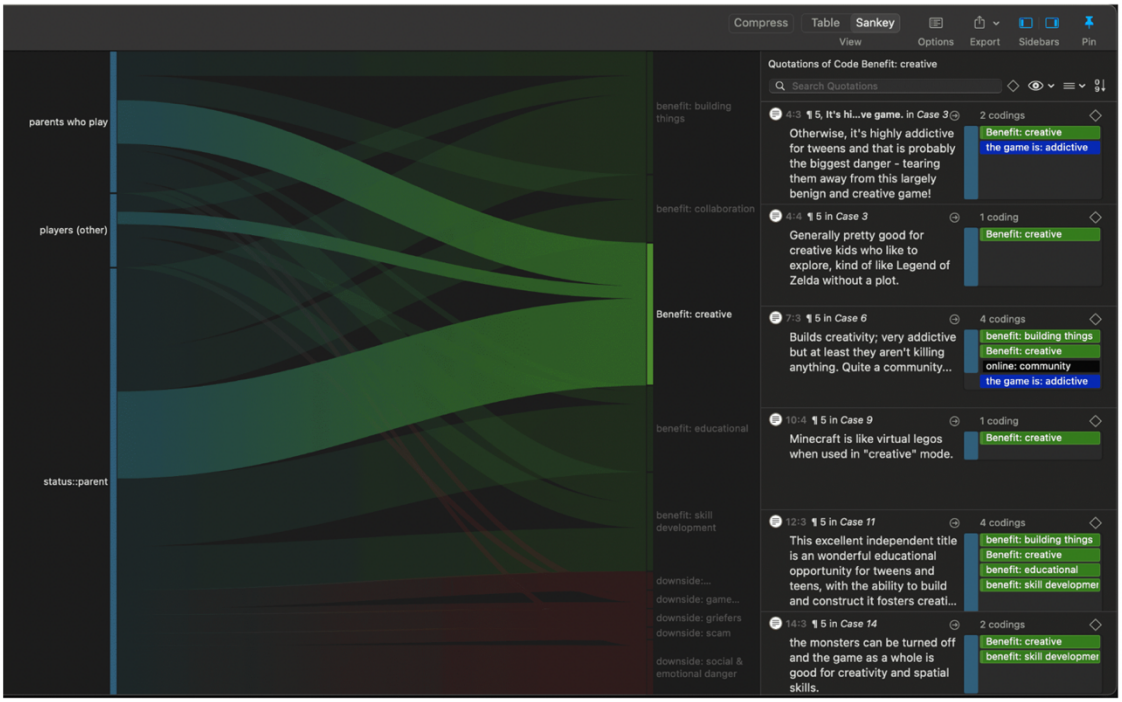


Figura 3. Finestra di ATLAS.ti© con una delle possibili modalità di visualizzazione dei dati codificati, nello specifico mediante diagramma di Sankey. Esso consente di evidenziare visivamente concetti complessi, concentrandosi su un singolo aspetto o risorsa che si desidera mettere in risalto; offre inoltre l'ulteriore vantaggio di supportare più livelli di visualizzazione facendo risaltare le componenti dominanti e le grandezze relative e/o le aree con i maggiori riscontri[14].

BIBLIOGRAFIA

[1] Bruner, Emiliano. *La mente oltre il cranio. Prospettive di archeologia cognitiva*. Roma: Carocci editore, 2018.

[2] Buonincontri, Piera, Giulia Caneva, Carla Maurano, e Maria Simeon. «Il patrimonio culturale materiale e immateriale». In *Il futuro dei territori antichi. Problemi, prospettive e questioni di governance dei paesaggi culturali evolutivi viventi*, (a cura di) Ferruccio Ferrigni, 35–40. Ravello: Centro Universitario Europeo per i Beni Culturali, 2013.

[3] Cortellazzo, Michele. «Metodi qualitativi e quantitativi di analisi dei testi». *Contemporanea* 16, fasc. 2 (giugno 2013): 299–310.

[4] Figuera, Marianna. *Past for the future: archeologia, conservazione e nuove tecnologie. Casi studio greci e italiani*. Roma: Edizioni Quasar, 2022.

[5] Hamilakis, Yannis, Mark Pluciennik, e Sarah Tarlow. *Thinking through the body: Archaeologies of Corporeality*. Berlin: Springer, 2012.

[6] Landeschi, Giacomo, e Eleanor Betts. *Capturing the Senses: Digital Methods for Sensory Archaeologies*. Berlin: Springer Nature, 2023. https://link.springer.com/book/10.1007/978-3-031-23133-9#toc.

[7] Marazzi, Antonio. *Antropologia dei sensi. Da Condillac alle neuroscienze*. Roma: Carocci editore, 2010.

[8] Milesi, Patrizia, e Patrizia Catellani. «L'analisi qualitativa dei testi con il programma Atlas.ti.» In *Metodi qualitativi in psicologia sociale*, (a cura di) Bruno Mazzara, 283–304. Roma: Carocci editore, 2002. https://www.researchgate.net/publication/248708093_L'analisi_qualitativa_di_testi_con_il_programma_Atlasti.

[9] Pink, Sarah. *The Future of Visual Anthropology. Engaging the senses*. London-New York: Routledge, 2006.

[10] Quevedo Diaz, Marcos. *Il cervello inconscio. Gli automatismi della nostra mente*. Milano: EMSE, 2022.

[11] Schutt, Russell K. «Qualitative Data Analysis». In *Investigating the Social World*, 320–57. London: Sage Publications, 2011.

[12] Skeates, Robin, e Jo Day. *The Routledge Handbook of Sensory Archaeology*. London: Routledge, 2022.

[13] Tilley, Christopher. *A Phenomenology of Landscape: Places, Paths and Monuments*. London: Berg Publishers, 1994.

[14] Townend, Steve. «Review of Muhr, T. 1997. ATLAS.ti 5: The Knowledge Workbench». *Papers from the Institute of Archaeology* 14 (2003): 161–69. https://www.researchgate.net/publication/307646114_Muhr_T_1997_ATLASti_5_The_Knowledge_Workbench.

[14] https://doc.atlasti.com/ManualMac/CodeDocumentTable/CodeDocumentTable.html (vd. nota 10).

Godscapes: Modeling Second Millennium BCE Polytheisms in the Eastern Mediterranean thanks to the Storage project

Nicola Laneri[1], Marianna Nicolosi-Asmundo[2], Daniele Francesco Santamaria[3], Chiara Pappalardo[4]
[1]Department of Humanities, University of Catania, Ital - nlaneri@unict.it
[2]Department of Mathematics and Computer Sciences, University of Catania, Italy - marianna.nicolosiasmundo@unict.it
[3]Department of Mathematics and Computer Sciences, University of Catania, Italy - daniele.santamaria@unict.it
[4]Department of Humanities, University of Catania, Italy - chiara.pappalardo@unict.it

ABSTRACT
The Godscapes project proposes to combine a material approach with the Semantic Web scientific model to investigate cultural transformation and, more specifically, how external elements trigger the transformation of religiosity, resulting in new hybrid elements. In fact, focusing as a case-study on the Levant during the second millennium BCE, a period marked by intense and long-reaching cultural exchanges, the project investigates the interplay between indigenous and exogenous elements (Egyptian, Syro-Mesopotamian, Aegean, Anatolian) in shaping polytheistic beliefs and practices through the analysis of four types of data – funerary, architectural, iconographic and textual. Thus, the project addresses a new scientific perspective emphasizing the use of material culture to understand the connection between humans and the divine, with a focus on unravelling past religious hybridization to grasp, in particular, how the second millennium cultural and religious intermingling persisted in the syncretic experience leading to the construction of the Israelite monolatry in the first millennium BCE.

KEYWORDS
Materiality; Semantic Web; Levant; Religion; Second Millennium BCE.

1. INTRODUCTION
The project 'Godscapes: Modeling Second millennium BCE Polytheism in the Eastern Mediterranean'[1] is a National Relevance Research Project (PRIN 2020) led by the University of Catania in collaboration with the University of Rome Sapienza, the University of Pisa, and the National Research Council (CNR ISPC). Main aim of the project is to define a new scientific method that will be applied to understand how humans entangle with the divine. To reach this target, it combines a material perspective that, following the most recent approaches to religious studies that have formed the so-called 'material turn', stresses the pivotal role of material culture in shaping beliefs and practices, with an innovative use of the artificial intelligence, namely, the Semantic Web.

Such an approach, which will prove beneficial in the understanding of both ancient and modern forms of religiosity, is tested on the complex forms of polytheisms that were practiced in the Levant during the second millennium BCE. In fact, the period chosen as case-study was characterized by intense international exchanges, both commercial and diplomatic, that from the North Africa and the Eastern Mediterranean reached as far as Central Asia, enhancing the circulation of people, things and ideas, and the creation of original forms of religious syncretism. Such a phenomenon is more evident in the areas where the contacts flourished and multiplied because of their centrality in the exchange process. Within such a perspective, the Levant, mainly referring to modern-day Israel, Palestine, Jordan, Lebanon and Syria, represented an extraordinary corridor not only for trade routes and military expansion, but also to people, customs and beliefs from the whole Eastern Mediterranean basin and beyond. It is in this complex cultural milieu, where local, exogenous and hybrid elements interacted with one another, that lie the roots of the Israelite monolatry of the first millennium BCE.

Targeting the phenomena of cultural hybridization that affected the religiosity of the communities inhabiting the Levant during the second millennium BCE in order to understand how the polytheistic sources had been rethought, triggered and reshaped in the process that will bring to the Israelite monotheism, Godscapes focuses on the analysis of four types of data – funerary, architectural, iconographic, and textual. The approach will: 1) identify the exogenous and endogenous layers of religiosity; 2) define the diagnostic markers of the second millennium BCE Levantine religiosity; 3) demonstrate how a syncretic outcome (i.e., the Israelite monotheism) can be considered as the result of a complex network of inter-religious encounters originated during the Second Millennium BCE.

[1] https://godscapes.unict.it/

2. STATE OF THE ART

The ongoing discourse concerning the roots of the Israelite monolatry has sparked intense debates throughout human history. Based on fascination, or on attempts to combine Egyptian and Biblical sources, scholars have speculated about an Egyptian origin, pinpointing the Late Bronze Age II, that witnessed the ascendancy of Amenhotep IV (1353-1336 BCE), the Pharaoh who ushered in an innovative henotheistic belief centred around the god Aten, as a crucial period [1]. However, holistic approaches considering the Israelite monolatry within the complex cultural milieu of the second millennium BCE Levantine polytheistic traditions were only rarely applied [10]. In fact, during this phase demanded raw materials and artifacts widely circulated throughout the Eastern Mediterranean and the north-eastern part of Africa, leading to a process of 'international' exchanges. The movement of 'materials' enhanced the movement of people, ideas, culture, and knowledge. Such a phenomenon is more evident – and more readable for modern scholars – in the areas where the contacts flourished and multiplied because of their centrality in the exchange process. Within such a perspective, the Levant, including modern Israel, Palestine, Jordan, Syria and Lebanon, represented an area of great resilience, absorbing, transforming and integrating external influences, ideas, practices and cultures.

Phenomena of cultural transfer and material entanglement can especially be envisioned in the remarkable consumption of Egyptian and Egyptianizing material culture from across the Levant. Egyptian-type temples, objects from funerary contexts, and cultic paraphernalia, such as the Qudshu plaque figurines, clay cobras, or the large quantity of royal scarabs associated with Amenhotep III and his wife Tiy, may point not only to ritual activity but also to the presence of Egyptian cultic personnel. Sites such as Sidon, Qatna, and Ugarit reveal how ideas and beliefs had passed onto the local society, and archaeological remains expressing forms of entanglement are even more explicit at Byblos: here, for instance, the Egyptian goddess Hathor is labelled as lady of Byblos; the god Reshep assumes Egyptian posture and iconography; figures of Anubis and Isis are scattered across the site. In fact, as correctly pointed out by Mark S. Smith [14], the archaeological research in the area has created a massive amount of data fundamental to interpret the relationship between biblical monotheism and polytheism in ancient Israel and neighboring cultures. However, research projects aiming at explaining how the interaction shaped local ritual and religion at both a public and private level have been carried out so far only at sparse sites [11].

3. METHODOLOGIES

The methodology presented aims to create a model that elucidates the relationships among different material elements of religiosity during cultural transformations, adopting a comprehensive approach that considers the entanglement of material elements, humans, and ideas in the construction of religious beliefs and practices [12]. In so doing, the research addresses fundamental questions about the transformation of autochthonous forms of religiosity by exogenous elements, identifying persisting and abandoned elements, and examining the social contexts in which these elements are most recognizable.

Focusing on a defined geographical region (i.e., the Levant) and period (i.e., the second millennium BCE), the method involves deconstructing four fundamental aspects associated with material religiosity: religious architecture, religious iconography, funerary rituals/beliefs, and religious texts. The work of deconstruction is operated through the application of an epistemological process in which a suite of semantic ontologies is designed, implemented, linked to the most popular semantic web vocabularies in the field, and populated through a process of data entry. For instance, information regarding a temple of the Second Millennium BCE (e.g., the *migdal* temple of Shechem) are entered in the dataset describing religious architecture according to specific criteria such as temple plan, orientation, and presence of semi-fix furniture that respond to a distinction between apparently exogenous (i.e., a western Syrian plan type) and indigenous (i.e., the use of outdoor sacred spaces) elements. In order to populate the web ontologies, the research project currently relies on data from a set of pilot archaeological sites: Ugarit, Byblos, Shechem, Megiddo, Jericho, Lachish, Deir el-Balah, and Serabit el-Khadim (see Fig. 1).

Such a deconstruction employs the Web Ontology Language 2 (OWL 2) to represent a formal, shared conceptualization of the domain responding to the FAIR principles, namely, findability, accessibility, interoperability, and reusability, which guide Semantic Web standards. The process of populating web ontologies with a vast amount of published data is crucial for enabling effective querying of interconnected knowledge. In the context of the current project, where data of varied nature is collected from different sources across a wide geographical area, relational databases face limitations in gathering, collecting, and integrating comparable information. The Semantic Web, with its well-established methodologies and tools, provides a solution to semantically model application domains, integrate data, and make it globally accessible [2]. The Semantic Web involves machine-readable data that enable software agents to query and manipulate information autonomously, promoting increased coherence and dissemination of knowledge. Automated reasoning procedures in the Semantic Web allow for the extraction and processing of implicit information, facilitating a deeper understanding of the domain [15].

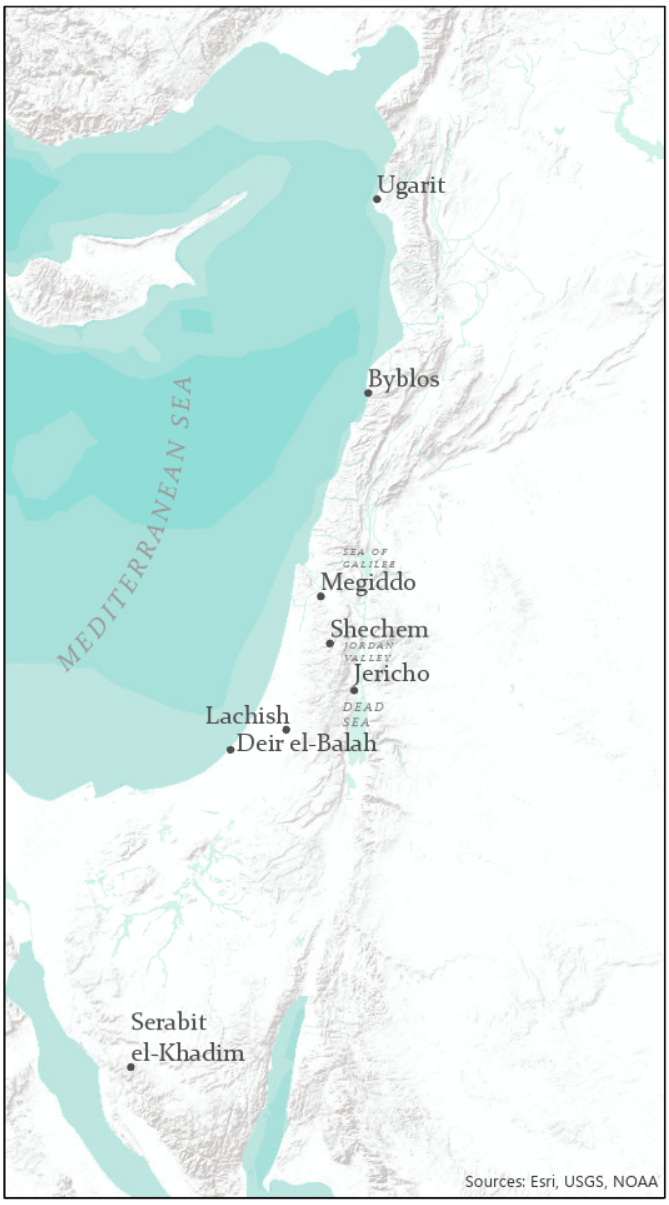


Figure 1. Map of the pilot archaeological sites whose data have been entered into the Godscapes database

Hence, thanks to the Semantic Web the four types of data collected so far will be interconnected, while reasoning and semantic query will discover which exogenous elements transformed the indigenous forms of religiosity, and in which context these elements are recognizable. For instance, through the semantic modelling of the already mentioned *migdal* temple of Shechem, the reasoners will distinguish the exogenous elements from the local ones in a way that cannot be achieved by adopting relational databases, whereas Large Language Models cannot be trained to answer the research questions without a sufficient number of samples and their correlation with the expected output.

An ontology, defining representational primitives such as classes and properties, is essential in this context. The W3C Web Ontology Language (OWL) serves as a Semantic Web language designed to represent intricate knowledge about things, groups of things, and relations between them. OWL, recognized as the standard for representing ontologies by the World Wide Web Consortium (W3C), provides users with constructs that enhance the design of ontologies in real-world domains beyond the capabilities of basic semantic web models like RDF and RDFS. RDF allows structured and semi-structured data to be shared across web applications, while RDFS extends it with taxonomies and a more expressive set of primitives for defining range, domain, and subsumption axioms. The possibility of constructing defined, complex OWL concepts paves the way for non-trivial, sometimes unexpected inferences. Moreover, introducing rule of the Semantic Web Rule Language (SWRL) allows one for definitions not expressible in OWL 2 and it is often convenient in terms of efficiency of the reasoners.

The structure of the four 'Godscapes' OWL ontologies is mainly inspired by ontological models developed in previous research works and by relevant standard ontologies in the field. Specifically, to model the religious architecture encountered in our current research on the second millennium BCE Levant, the ontologies presented in Cantale et al. 2017 [6], and in Cantale et al. 2021 [5] are considered. The first one describes the relevant structural components of Catania's Benedectine Monastery "San Nicolò l'Arena" while the second one proposes a definition of the architectural type of monastery. Regarding objects, we model our ontologies by extending the Ontoceramic project [3].

A similar approach is used for the characterization of religious iconography and funerary data (see Fig. 2), where the EPIONT taxonomy, an ontology for epigraphs implemented by specializing a portion of the CIDOC CRM model, is considered [7]. Finally, for the definition of an ontology concerning religious texts, we will take into account CRMtex, an extension of CIDOC CRM ontology formally representing the specific requirements of the studies in ancient texts, including papyrology, paleography, codicology and epigraphy [9]. While CIDOC is the actual standard for integrating cultural heritage data, other ontologies such as the ones provided by the Plaides project for the historical toponymy, will be useful. The TGO ('The Godscapes Ontology', see Fig. 3) will finally be endowed with a set of ontological primitives to connect the four ontologies mentioned above, and to encompass the huge amount of digital and non-digital bibliography related to Godscapes.

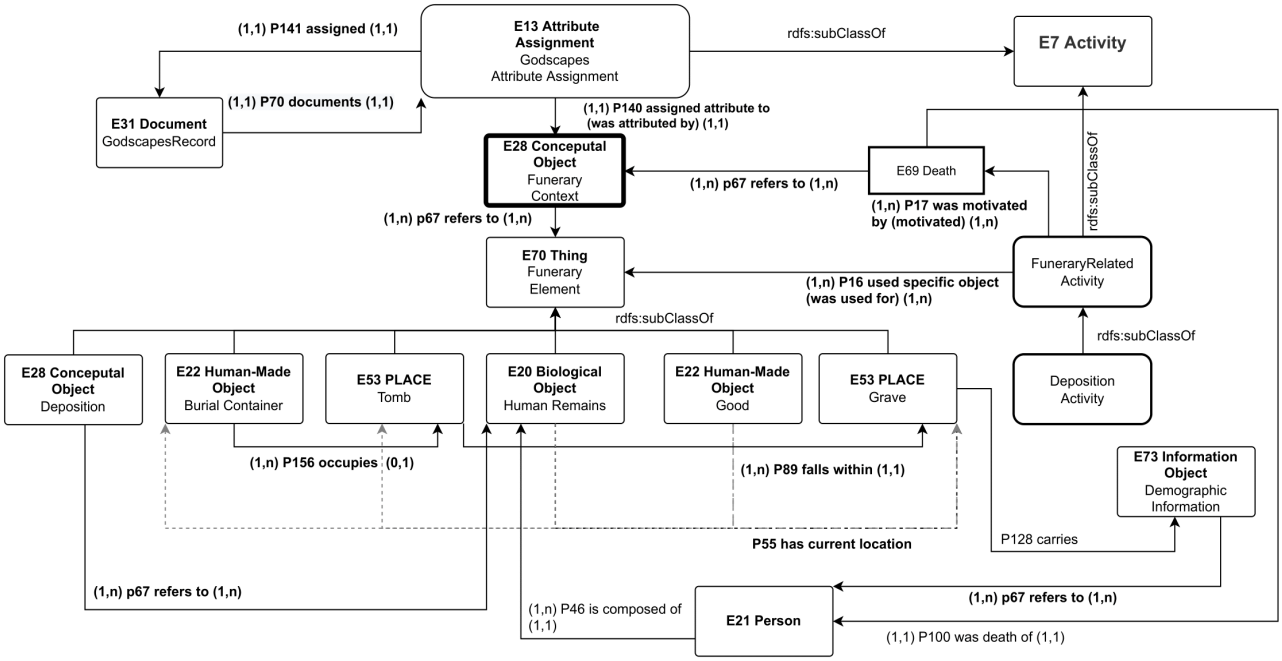


Figure 2. Model design of the Godscapes ontology for the funerary data

4. OBJECTIVES AND EXPECTED RESULTS

Based on these methodological premises, the aims of the Godscapes project are three folded:

1. Create a coherent model that defines basic elements of material religiosity through a detailed analysis of architectural, iconographic, funerary and textual data in a specific and coherent geographical and chronological frame. Not only such a model will be useful for the definition of forms of material religiosity within ancient Near Eastern societies, but it will furnish a tool to be implemented also with other datasets.
2. Interpret the transformation of a polytheistic religious phenomenon through a clear and diachronic definition of how exogenous and indigenous elements, verifiable in the material remains of religious practices and beliefs, interacted in the process of forming the premises of a completely different form of religiosity as it is the Israelite monolatry. At this stage a suite of SWRL rules is defined, describing interactions between relevant elements of material religiosity.
3. Make the project datasets open to the public through the use of 'open science practices.' 'As open as possible, as closed as necessary' is the statement incorporated in the missions of the newly funded Horizon Europe program and that are among the fundamental pillars of the 2030 Agenda for Sustainable Development. Open data (i.e., free from copyright and shareable as public domain) are the future of research and, consequently, open science practices will increase the quality and impact of Responsible Research & Innovation (RR&I), leading to great responsiveness to societal challenges through, for example, the European Open Science Cloud (EOSC).

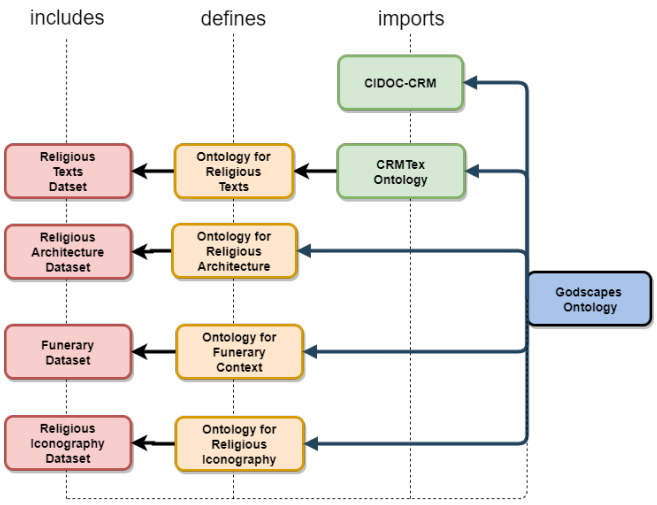


Figure 3. Diagram for TGO

Following these objectives, the anticipated outcomes of the Godscapes research project include the creation of a suite of OWL 2 ontologies based on the vocabularies of religious materiality, and the implementation of an openly accessible database containing all pertinent datasets utilized in the project. The dataset, available in open format, will be accompanied by formalized SWRL rules and queries, in such a way as to characterize specific concepts on the one side and answer research questions from the other.

To enhance accessibility, a web platform and a WebGIS interface will be developed to query and visualize the data in a user-friendly manner. This achievement aligns with the best practice for a NextGen of scientific instrumentation, tools and methods for advanced digital solutions. pursues the creation of network connectivity and enabling collaboration without boundaries. These are both Destinations of the Research Infrastructures (RI) forecasted in the newly planned Horizon Europe program that aims at creating interconnected services providing fair and quality certified Open Data.

5. ACKNOWLEDGEMENTS

The Godscapes project acknowledges the Italian Ministry for University and Research (PRIN 2020, SH6-Classical Antiquity), and the Storage project (University of Catania, program 'PIA.CE.RI.') for their invaluable support and funding.

REFERENCES

[1] Assmann, Jan. *Moses Der Ägypter: Entzifferung Einer Gedächtnisspur*. München: Hanser, 1999.

[2] Berners-Lee, Tim, James Hendler, and Ora Lassila. 'The Semantic Web'. *Scientific American* 284, no. 5 (2001): 34–43.

[3] Brancato, Rodolfo, Marianna Nicolosi-Asmundo, Grazia Pagano, Daniele F. Santamaria, and Salvatore Ucchino. 'An Ontology for Legacy Data on Ancient Ceramics of the Plain of Catania'. In *Proceedings of the 34th Italian Conference on Computational Logic, Trieste, Italy, June 19-21, 2019*, edited by Alberto Casagrande and Eugenio G. Omodeo, 2396:59–67. CEUR Workshop Proceedings, 2019.

[4] Brancato, Rodolfo, Marianna Nicolosi-Asmundo, Grazia Pagano, Daniele F. Santamaria, and Salvatore Ucchino. 'Towards an Ontology for Investigating on Archeological Sicilian Landscapes'. In *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage Co- Located with the 31st International Conference on Advanced Information Systems Engineering, ODOCH@CAiSE 2019. Rome, Italy, June 3*, edited by Antonella Poggi, 2375:85–90. 2019: CEUR Workshop Proceedings, 2019.

[5] Cantale, Claudia, Domenico Cantone, Manuela Lupica Rinato, Marianna Nicolosi-Asmundo, Daniele Francesco Santamaria, and Maria Rosaria Stufano Melone. 'The Ideal Benedictine Monastery: From the Saint Gall Map to Ontologies'. *Applied Ontology* 16, no. 2 (2021): 137–60.

[6] Cantale, Claudia, Domenico Cantone, Marianna Nicolosi-Asmundo, and Santamaria, Daniele Francesco. 'Distant Reading Through Ontologies: The Case Study of Catania's Benedictines Monastery'. *Italian Journal of Library, Archives and Information Science* 8, no. 3 (2017): 203–19.

[7] Cantone, Domenico, Salvatore Cristofaro, Marianna Nicolosi-Asmundo, Francesca Prado, Daniele Francesco Santamaria, and Daria Spampinato. 'An EpiDoc Ontological Perspective: The Epigraphs of the Castello Ursino Civic Museum of Catania via CIDOC CRM'. *Archeologia e Calcolatori* 30 (2019): 139–57.

[8] Cantone, Domenico, Marianna Nicolosi-Asmundo, Daniele F. Santamaria, and Francesca Trapani. 'OntoCeramic: An OWL Ontology for Ceramics Classification'. In *Proceedings of the 30th Italian Conference on Computational Logic, Genova, Italy, July 1-3, 2015*, edited by Davide Ancona, Marco Maratea, and Viviana Mascardi, 1459:122–27. CEUR Workshop Proceedings, 2015.

[9] Felicetti, Achille, and Francesca Murano. 'Scripta Manent: A CIDOC CRM Semiotic Reading of Ancient Texts'. *International Journal on Digital Libraries* 18, no. 4 (2017): 263–70.

[10] Killebrew, Ann E. *Biblical Peoples and Ethnicity: An Archaeological Study of Egyptians, Canaanite*. Vol. 9. Society of Biblical Lit, 2012.

[11] Koch, Ido. 'Religion at Lachish under Egyptian Colonialism'. *Die Welt Des Orients* 49, no. 2 (2019): 161–82.

[12] Laneri, Nicola. *From Ritual to God in the Ancient Near East. Tracing the Origins of Religion*. Cambridge: Cambridge University Press, 2024.

[13] Oggiano, Ida. *Dal terreno al divino: archeologia del culto nella Palestina del primo millennio*. Vol. 20. Roma: Carocci, 2005.

[14] Smith, Mark S. *The Origins of Biblical Monotheism: Israel's Polytheistic Background and the Ugaritic Texts*. Oxford: Oxford University Press, 2001.

[15] Staab, Steffen, and Rudi Studer. *Handbook on Ontologies*. Springer Science & Business Media, 2010.

Il progetto Storage: dai dati al Web

Simone Faro[1], Pietro Maria Militello[2], Marianna Nicolosi-Asmundo[3]
[1] Department of Mathematics and Computer Science, University of Catania, Italia - simone.faro@unict.it
[2] Department of Humanities, University of Catania, Italia - milipi@unict.it
[3] Department of Mathematics and Computer Science, University of Catania, Italia - marianna.nicolosiasmundo@unict.it

ABSTRACT
L'intervento presenta i risultati di un progetto di ricerca finanziato dall'Ateneo di Catania, il cui obiettivo è la valorizzazione dei *Legacy Data* in ambito archeologico e storico-artistico e la elaborazione di modelli di fruizione adeguata anche sul web.

PAROLE CHIAVE
Bene culturale; archivi digitali; ontologie; integrazione testuale.

1. INTRODUZIONE
La digitalizzazione dei Beni Cultuali è precondizione essenziale per una moderna gestione, conservazione e fruizione pianificata, questa digitalizzazione non presenta solo difficoltà operative, l'inserimento di grandi quantità di dati, ma anche nodi teorico-scientifici, e nodi organizzativi. Nonostante gli sforzi degli specialisti di organizzare la complessità della cultura materiale attraverso la creazione di tassonomie e di ontologie specifiche[1], è evidente l'estrema eterogeneità che caratterizza i dati relativi al Patrimonio Culturale, non solo tra classi molto diverse tra di loro di beni (libri, documenti di archivio, monumenti, manufatti, aree geografiche, ecc.) ma anche all'interno dei singoli settori (si pensi, per esempio, al problema della tassonomia ceramica).
Il nodo teorico è dunque duplice: per un verso si tratta di compiere scelte di classificazione e rappresentazione del Patrimonio Culturale (comprensivo sia di "testi" che di "oggetti"), che facciano tesoro dello stato dell'arte; dall'altro di integrare i cosiddetti *legacy data*, termine con cui si fa riferimento ai dati di vecchio stampo ed al loro valore di "eredità" per la ricerca ma anche per la tutela. Per un altro verso, si tratta di rendere queste informazioni accessibili in maniera interfacciabile con i grandi *repositories*, e insieme *user -friendly*, per consentire un utilizzo ampio a diverse tipologie di soggetti interessati. Il problema organizzativo è quello di tenere insieme il nodo operativo (necessità di immettere grandi quantità di dati) e quello teorico (procedere in modi scientificamente avanzati).
Nel processo, concetto chiave non è solo la normalizzazione del dato, ma anche quello del suo possibile riuso, in linea con i *FAIR pinciples* (*Findable*, *Accessible*, *Interopeable*, *Re-usable*) [15]. Il riuso è certamente un aspetto complesso, che in archeologia coinvolge non solo l'integrità del dato, la sua provenienza, trasparenza e riproducibilità, ma anche la sua interpretazione [12].
Questi aspetti sono stati alla base della ricerca del progetto "Storage. Dai dati al Web", cui hanno contribuito umanisti e informatici, e che affronta il problema della raccolta, archiviazione, gestione e comunicazione digitale dei dati in ambito archeologico e storico-artistico, sia sul campo che in ambito museale e archivistico. Al progetto hanno partecipato docenti del DISUM e del DMI dell'Università di Catania[2]. Obiettivo del progetto Storage, di cui presentiamo qui alcuni risultati preliminari, sono: 1) l'avanzamento metodologico nella raccolta, ricostruzione, gestione e condivisione di oggetti digitali, secondo standard adeguati alla normativa nazionale, garantendo interoperabilità tra i diversi sistemi; 2) lo sviluppo di soluzioni algoritmiche per la MTR in ambito archeologico; 3) la creazione di un'ontologia che modelli adeguatamente il Patrimonio Culturale, i relativi strumenti di derivazione e interrogazione; 4) l'accessibilità in rete di grandi quantità di dati, utilizzando alcuni casi studio ad ampio spettro (archivi, manufatti artistici, manufatti archeologici).

2. LA GESTIONE E CONDIVISIONE DI ARCHIVI DIGITALI
Per quanto riguarda il primo punto, raccolta, ricostruzione, gestione e condivisione di oggetti digitali, sono stati scelti casi studio appositamente diversificati: siti archeologici di area siciliana, albanese e dello Azerbaijan; collezioni archeologiche (Museo di Archeologia del DISUM dell'Università di Catania; Museo Stratigrafico di Festòs a Creta), collezioni didattiche (Archivio dell'ex Istituto di Archeologia del DISUM) e collezioni storico-artistiche (Collezioni del Museo di Castello Ursino a Catania). Le attività hanno riguardato sia la gestione dei dati esistenti, sia la creazione di dati nativi digitali attraverso scansioni laser o fotogrammetria. In questo contributo ci focalizziamo sul primo aspetto, lasciando il secondo ad

[1] Una su tutte l'ontologia del Cultural Heritage CIDOC-CRM (http://cidoc-crm.org/) [8, 9].

altro contributo.
Un esempio dei problemi posti nella Introduzione è costituito dalla documentazione relativa agli archivi del DISUM, Archivio dell'ex Istituto di Archeologia [1, 16], Archivio della Missione Archeologica di Festòs [14], Museo di Archeologia [18], che custodiscono diversi materiali provenienti dai vari gabinetti ed istituti di Archeologia che si sono susseguiti presso l'Ateneo catanese fino al 1999, anno della istituzione dei dipartimenti. Il materiale comprende strumentazioni per la riproduzione audiovisiva, materiale didattico, tesi di laurea, e soprattutto una grande quantità di materiale fotografico comprensivo di diapositive, negativi, microfiches, e una serie di taccuini e relazioni di scavo. A questi si aggiungevano pezzi delle collezioni archeologiche (ceramica, litica ecc.). Parte di questo materiale era stato catalogato, tuttavia, in modo prevalentemente cartaceo, una parte digitalizzata negli anni '90, sotto forma di database in formato .db3. Solo per alcuni casi è stato possibile costruire un modello di database ad hoc in grado di soddisfare tutte le esigenze della ricerca [11, 12], nella maggior parte dei casi si è invece dovuto tenere conto della normativa esistente. Questa fa frequentemente (ma non sempre) riferimento alle schede dell'Istituto Centrale per il Catalogo e la Documentazione[2]. Mentre la fruizione dei dati ivi custoditi è libera (ma con i limiti sotto indicati) l'accesso alla compilazione può avvenire solo dopo il riconoscimento da parte dell'ICCD dello statuto di Ente Schedatore, statuto che può valere tuttavia solo per la catalogazione di beni di proprietà dell'Ente stesso (in questo caso l'Università), ma necessita, negli altri casi, di autorizzazioni specifiche da parte degli enti proprietari, con un meccanismo burocratico non sempre fluido. Aggiungiamo che le schede dell'ICCD, create per rispondere al numero più ampio di contesti possibili, risultano spesso ridondanti per i casi specifici. Infine, la consultazione delle schede sul sito nazionale è libera, ma organizzata per classi di materiale e non consente un agevole incrocio delle informazioni appartenenti a due categorie diverse di dati (per es., fotografie e reperti archeologici), laddove un obiettivo del progetto Storage era l'utilizzo di un formato di catalogazione che consentisse una interrogazione dei dati capace di attivare ulteriori livelli di informazione oltre quelli richiesti.
Verificata la impossibilità di costruire una struttura ad hoc, per i limiti sopra citati, le soluzioni hanno dovuto adattarsi alle diverse situazioni. In un caso (archivio di Festòs) esisteva una documentazione digitale risalente alla fine degli anni '80, realizzata prima in .db3 per quei tempi innovativi, e riversata successivamente in Access. Per questa si è progettata una riorganizzazione del disegno dati atta al riversamento nel database di tipo relazionale che è stato creato da M. Figuera e E. Platania, sempre nell'ambito del progetto "Storage". Quest'ultimo, anche se realizzato per la gestione dei reperti cd. "minori", si presta anche al trattamento di dati di tipo archivistico per l'attenzione posta nella sua struttura agli aspetti documentali (molteplici fonti, numeri di inventario, collocazione, immagini, ecc.).
La catalogazione dei materiali è avvenuta partendo dalla tipologia di informazione delle schede dell'ICCD, con attenzione particolare a quel tipo di informazione "contestuale" (anno di acquisizione, ordinante) che consentisse la correlazione tra, per es., l'acquisto di una serie fotografica e l'attività didattica svolta in un determinato periodo. Sono state in questo modo schedate 5484 schede di materiali, 3500 lastre fotografiche, 299 negativi, 33 tesi di laurea, diverso materiale didattico e una quarantina di strumenti di rilievo e di riproduzione audiovisiva.

3. SVILUPPO DI SOLUZIONI ALGORITMICHE PER LA MTR IN AMBITO ARCHEOLOGICO

Uno dei problemi che emerge nel trattamento dei dati legacy riguarda l'incompletezza testuale di documenti, sia su supporto cartaceo che in altri formati, danneggiati dal trascorrere del tempo. Questa incompletezza si manifesta in annotazioni su taccuini, vecchie schede di inventario, descrizioni e simili. Un ambito prioritario del nostro progetto si è concentrato sull'affrontare il problema della "ricostruzione di frammenti mancanti" (Missing Text Reconstruction o MTR)all'interno di testi di natura epigrafica o archivistica [17]. Esso rappresenta una complessa sfida che emergequando manoscritti storici o culturali presentano parti illeggibili o assenti. Queste lacune possono derivare da diversi fattori, quali l'usura del tempo, danni fisici o lacune volute, come censura o danneggiamenti accidentali. L'obiettivo è restaurare il contenuto originale del testo attraverso la ricostruzione o il completamento delle parti mancanti.
Il processo di integrazione delle porzioni mancanti spesso implica l'utilizzo di metodologie informatiche avanzate [6, 10], in particolare algoritmi di elaborazione del linguaggio naturale e tecniche di text processing [18]. Questi approcci mirano a identificare pattern linguistici, parole o frasi coerenti con il contesto circostante e adatti al linguaggio e allo stile del testo originale.
Nel contesto specifico dei manoscritti antichi, le sfide sono numerose. La variabilità linguistica nel corso del tempo, le differenze stilistiche e la presenza di espressioni o parole obsolete complicano la ricostruzione delle porzioni mancanti. Inoltre, la mancanza di contesto circostante aumenta l'incertezza nella selezione delle parole o frasi corrette.

[2] ICCD, si veda il sito: http://www.iccd.beniculturali.it/it/1/home

Da un punto di vista pratico, risolvere il problema dell'integrazione di porzioni mancanti richiede una sinergia tra competenze umanistiche e informatiche. Gli esperti in storia, linguistica e cultura forniscono contesto e conoscenze linguistiche, mentre specialisti informatici sviluppano e applicano algoritmi avanzati per migliorare l'accuratezza del processo di integrazione. Le metodologie proposte nel progetto impiegano algoritmi basati su tecniche avanzate di string matching [10] e algoritmi combinatori [18], che svolgono un ruolo fondamentale nell'identificare con precisione i frammenti mancantiall'interno dei manoscritti. Basandosi sulla struttura della lingua, essi analizzano pattern di parole, lunghezza delle frasi ealtri elementi linguistici per proporre completamenti possibili. Un approccio avanzato prevede la creazione di modelli linguistici probabilistici, integrando il contesto storico e culturale del manoscritto. Analizzando varie combinazioni di parole e frasi, gli algoritmi valutano la coerenza con il contesto circostante e la grammatica della lingua. L'implementazione di tecniche di pruning per eliminare soluzioni non plausibili contribuisce a ridurre il tempo computazionale delle proposte, migliorando l'accuratezza complessiva del processo.
Gli approcci proposti si basano sulla costruzione di automi a stati finiti deterministici che contengono informazioni statistiche sulla struttura del testo. Questi automi vengono utilizzati per la predizione delle porzioni da ricostruire e sono ottenuti da porzioni integre di testo nella stessa lingua. Pertanto, l'integrazione di testi mancanti presenta diverse sfide.
I primi test sperimentali dell'approccio proposto sono stati condotti su testi artificiali, nei quali sono state deliberatamente introdotte lacune di varie dimensioni al fine di simulare situazioni realistiche. I risultati ottenuti hanno evidenziato l'abilità del sistema nel ricostruire con notevole precisione la porzione di testo mancante, a condizione che un corpus significativo sia disponibile per l'addestramento dell'automa. I risultati più promettenti emergono quando entrambi i lati, destro e sinistro, della porzione di testo mancante sono presi in considerazione dall'automa responsabile della ricostruzione.
È stato tuttavia evidenziato che la complessità della lingua e la variabilità stilistica rendono difficile la creazione di modelli linguistici affidabili. Inoltre, la presenza di lacune significative e la mancanza di contesto circostante possono limitare l'accuratezza della ricostruzione. La gestione dell'incertezza, la preservazione dell'autenticità del testo originale e la minimizzazione degli errori sono problematiche cruciali.
L'integrazione di testi mancanti in manoscritti antichi richiede un approccio equilibrato tra tecniche avanzate di string matching e algoritmi combinatori. Nonostante le sfide, l'applicazione di tali soluzioni offre notevoli vantaggi nella preservazione e comprensione del Patrimonio Culturale.

4. LA DEFINIZIONE DI UN'ONTOLOGIA PER LA RAPPRESENTAZIONE DEL PATRIMONIO CULTURALE E RELATIVI STRUMENTI DI INFERENZA E INTERROGAZIONE

Le ontologie computazionali sono mezzi per modellare la struttura di un sistema catturando le entità e le relazioni rilevanti, che emergono dalla sua osservazione e che sono utili per scopi ben precisi. Esse sono definite come specifiche formali ed esplicite di concettualizzazioni condivise, cioè di viste astratte e semplificate del mondo che, per qualche motivo, desideriamo rappresentare [17]. Le ontologie sono anche uno strumento standard del W3C per il SemanticWeb [13]. In tale ambito, il loro potenziale risulta particolarmente ricco in quanto, in combinazione con i Linked Data, le ontologie permettono non soltanto di definire rappresentazioni arbitrariamente astratte di un dominio d'interesse, mettendo in relazione a livello globale tassonomie, vocabolari e dati di natura eterogenea, ma anche di effettuare efficientemente l'interrogazione di conoscenza interconnessa. Inoltre, la presenza di opportuni strumenti di ragionamento, quali ad esempio i reasoner Hermit e Pellet, permette di verificare la consistenza di vocabolari e ontologieinferendo, ove possibile, nuova informazione.
Nell'ambito del progetto presentato in questo contributo, che riguarda la gestione di entità di diversa natura quali manufatti archeologici, documenti d'archivio e materiale fotografico, e di renderli globalmente fruibili mettendoli in relazione con altre risorse provenienti da altre sorgenti, la creazione di un'ontologia web che li modelli risulta essenziale.Il linguaggio di rappresentazione più ampiamente utilizzato per la costruzione di ontologie web è il Web Ontology Language (OWL). OWL, riconosciuto come standard W3C per rappresentare ontologie, fornisce agli utenti costrutti chepermettono di raggiungere un buon livello di espressività, rispetto ad altri linguaggi di rappresentazione per il Semantic Web quali RDF ed RDFSchema. In particolare, RDF permette ai dati strutturati e semi-strutturati di essere condivisi globalmente sul web, mentre RDFSchema estende RDF con l'introduzione di tassonomie, di primitive per la definizionedi dominio e range di una relazione ed assiomi di sussunzione.
La possibilità offerta da OWL di definire concetti complessi apre la strada ad inferenze logiche non banali e a volte inattese. Inoltre, l'aggiunta al modello ontologico di opportune regole del Semantic Web Rule Language (SWRL) permette allo sviluppatore di definire entità (concetti e relazioni) non esprimibili in OWL 2 (Web Ontology Language 2) e quindi di sviluppare un modello ontologico ancora più ricco e raffinato. Talvolta, l'uso delle regole SWRL al posto della definizione

di complesseespressioni OWL 2 è preferibile in termini di efficienza dei ragionatori. In questo progetto, la costruzione dell'ontologia viene portata avanti facendo riferimento ad ontologie già esistenti 'concettualmente vicine' e ad ontologie standard come ad esempio CIDOC CRM. In particolare, per la modellazione dei concetti più importanti riguardanti i manufatti archeologici usiamo ed estendiamo un'ontologia OWL 2 per la catalogazione e classificazione della ceramica antica chiamata OntoCeramic [7]. Essa è stata progettata con lo scopo di risolvere efficientemente problemi significativi quali la classificazione della ceramica rispetto alla forma, al tipo e alla classe, nonché l'analisi dei reperti rispetto alle loro componenti e ai luoghi di ritrovamento. L'ontologia è stata realizzata seguendo le schede di catalogazione standard ICCD. Successivamente, OntoCeramic è stata raffinata e arricchita migliorando le definizioni di (a) classe e tipo della ceramica, che aiuta a determinare il sito di produzione dei reperti archeologici, (b) la forma delle parti componenti, che aiuta a determinare il tipo di ceramica, e (c) le dimensioni della ceramica [3, 4]. OntoCeramic è stata definita secondo lo standard CIDOC CRM e utilizza l'ontologia LinkedGeoData per descrivere le località e per identificare il luogo di scoperta dei reperti archeologici. In questo progetto OntoCeramic viene estesa in modo tale da raffinare la classificazione della cronologia e supportare la gestione degli scavi stratigrafici, delle fabbrichedi produzione, dell'informazione topografica e dei riferimenti bibliografici. La classificazione della cronologia viene invece sviluppata ispirandoci al modello ontologico CoMOntology definito a partire dalla tavola cronologica di C. Broodbank dalla quale sono state individuate le classi principali: AbsoluteChronology, RelativeChronology, Area e ArcheologicalEvidence [2].
Per quanto riguarda la descrizione e modellazione dei documenti d'archivio e materiale fotografico viene usata ed estesa l'ontologia ArchivioMuseoDellaFabbrica, presentata in [5].
La coerenza del modello creato viene verificata utilizzando i reasoner Hermit e Pellet. Il modello ontologico viene poi popolato con i dati dell'archivio. Nuovamente, i ragionatori vengono eseguiti per determinare ulteriori deduzioni sui dati. Infine il dataset, disponibile in formato open, viene dotato di opportune regole SWRL e di competency questions SPARQL, linguaggio di interrogazione per il Semantic Web. Le prime sono idonee a caratterizzare concetti complessi specifici, mentre le seconde permettono di rispondere alle più comuni interrogazioni sul dataset.

5. CONCLUSIONI

Un primo esito del progetto Storage è stata l'interazione stretta, di studiosi del mondo umanistico e informatico su problemi concreti del Patrimonio Culturale. Un risultato è stato una maggiore consapevolezza degli operatori di area umanistica nella selezione degli strumenti computazionali finalizzata ad ottenere risultati non raggiungibili con gli strumenti tradizionali. Un uso del digitale, insomma, che non sia solo trasposizione di vecchi metodi in nuova veste. L'applicazione della *missing text reconstruction* è un caso esemplare. Particolarmente feconda è stata la discussione sullapossibilità di rispondere a quesiti che mettano in correlazione dati riversati su banche differenti, ma all'interno della stessapiattaforma, e sull'approfondimento delle classificazioni ontologiche del bene culturale e sulla applicazione dell'integrazione testuale.
Per quanto riguarda la classificazione, l'uso di formati semplificati rispetto alle normative proposte dall'ICCD non impedisce il riversamento dei dati ottenuti nel catalogo nazionale, e nello stesso tempo consente il potenziamento della comunicazione on line. Inevitabilmente si è rinunciato alla creazione di una unica banca dati, ma si è cercato a) di uniformare, quando possibile, quelle esistenti; b) di ovviare alla presenza di banche dati differenti utilizzando una unica piattaforma su cui riversare i dati, una volta ottenuta l'autorizzazione (come la piattaforma xDAMS, disegnata specificamente per i patrimoni archivistici). L'obiettivo finale è però più ambizioso, ed è quello di trasferire i dati sul web. Oggi le nuove modalità di condivisione dei dati sul web offerte dal paradigma dei *Linked Open Data* e dall'organizzazione semantica della conoscenza in ontologie computazionali, che consentono di strutturare la conoscenza relativa a un determinato dominio sulla base del significato dei termini utilizzati, consentirebbero agli Enti preposti alla gestione e valorizzazione del *Cultural Heritage* di contribuire attivamente al progetto di condivisione, che è potenzialmente significativo su molteplici livelli (scientifico, educativo, etico).

BIBLIOGRAFIA

[1] Bramante, Daniela M. «La classificazione e il riordino dell'archivio fotografico dell'ex Istituto di archeologia». In *Interferenze. Un dialogo tra scienze umane e scienze dure*, (a cura di) Figuera Marianna, 137–45. Catania, 2016.

[2] Brancato, Rodolfo, Marianna Figuera, Marianna Nicolosi-Asmundo, Daniele F. Santamaria, Paola Venuti, e Giuseppe Zappalà. «CoMOntology. Towards An Ontology for the Chronology of Mediterranean archaeologies: a model for the digital memory». In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 322-326, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[3] Brancato, Rodolfo, Marianna Nicolosi-Asmundo, Grazia Pagano, Daniele F. Santamaria, e Salvatore Ucchino. «An ontology for legacy data on ancient ceramics of the Plain of Catania». In *Proceedings of the 34th Italian Conference on Computational Logic,*

Trieste, Italy, June 19-21, 2019, (a cura di) Alberto Casagrande e Eugenio G. Omodeo, 2396:59–67. CEUR Workshop Proceedings, 2019.

[4] Brancato, Rodolfo, Marianna Nicolosi-Asmundo, Grazia Pagano, Daniele F. Santamaria, e Salvatore Ucchino. «Towards an ontology for investigating on archeological Sicilian landscapes». In *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage co- located with the 31st International Conference on Advanced Information Systems Engineering, ODOCH@CAiSE 2019. Rome, Italy, June 3*, (a cura di) Antonella Poggi, 2375:85–90. 2019: CEUR Workshop Proceedings, 2019.

[5] Cantale, Claudia, Domenico Cantone, Marianna Nicolosi-Asmundo, e Santamaria, Daniele Francesco. «Distant Reading Through Ontologies: The Case Study of Catania's Benedictines Monastery». *Italian Journal of Library, Archives and Information Science* 8, fasc. 3 (2017): 203–19.

[6] Cantone, Domenico, Simone Faro, e Oguzhan M. Külekci. «Shape-Preserving Pattern Matching». In *Proceedings of the Italian Conference on Theoretical Computer Science (ICTCS 2020)*, 2756:137–48. CEUR Workshop Proceedings, 2020.

[7] Cantone, Domenico, Marianna Nicolosi-Asmundo, Daniele F. Santamaria, e Francesca Trapani. «OntoCeramic: an OWL ontology for ceramics classification». In *Proceedings of the 30th Italian Conference on Computational Logic, Genova, Italy, July 1-3, 2015*, (a cura di) Davide Ancona, Marco Maratea, e Viviana Mascardi, 1459:122–27. CEUR Workshop Proceedings, 2015.

[8] Doerr, Martin. «Ontologies for Cultural Heritage». In *Handbook on Ontologies, International Handbooks on Information Systems*, (a cura di) Steffen Staab e Rudi Studer, 463–86. Berlin-Heidelberg: Springer, 2009.

[9] Doerr, Martin. «The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata». *AI Magazine* 24, fasc. 3 (settembre 2003): 75–92. https://doi.org/10.1609/aimag.v24i3.1720.

[10] Faro, Simone, e Stefano Scafiti. «Compact suffix automata representations for searching long patterns». *Theoretical Computer Science. Elsevier* 940 (2023): 254–68.

[11] Figuera, Marianna. «Database Management e dati archeologici: standardizzazione e applicazione della Logica Fuzzy alla gestione delle fonti e delle attribuzioni tipologiche». *Archeologia e Calcolatori* 29 (2018): 143–60.

[12] Figuera, Marianna. *Un sistema per la gestione dell'affidabilità e dell'interpretazione dei dati archeologici. Percezione e potenzialità degli small finds: il caso studio di Festòs e Haghia Triada*. Vol. 8. Praehistorica Mediterranea. Oxford: Archaeopress, 2020.

[13] Gruber, Thomas R. «Toward Principles for the Design of Ontologies Used for Knowledge Sharing». *International Journal of Human Computer Studies* 43, fasc. 5–6 (1995): 907–28.

[14] La Rosa, Vincenzo. *Radamante al Computer*. Catania, 2008.

[15] Martone, Maryann. «FORCE11: building the future for research communications and e-scholarship». *Bioscience* 65, fasc. 7 (2015): 635.

[16] Militello, Pietro M. «Immagini e strumenti. L'archeologia catanese attraverso l'archivio fotografico». In *Interferenze. Un dialogo tra scienze umane e scienze dure*, (a cura di) Marianna Figuera, 133–36. Catania, 2016.

[17] Studer, Rudi, Richard V. Benjamins, e Dieter Fensel. «Knowledge engineering: Principles and methods». *Data & Knowledge Engineering* 25, fasc. 1–2 (1998): 161–98.

[18] Tortorici, Edoardo. *La collezione Libertini e il Museo di Archeologia*. Catania, 2015.

La modellazione 3D al servizio dell'archeologia: nuove prospettive per l'applicazione ad edifici multipiano di età protostorica

Dario Puglisi[1], Marco Chiricallo[2]
[1] Università di Catania, Italia - dario.puglisi@unict.it
[2] Università degli Studi di Bari "Aldo Moro", Italia - marco.chiricallo@uniba.it

ABSTRACT[1]

Il contributo si propone di illustrare le nuove prospettive di ricerca aperte dall'applicazione della modellazione 3D all'analisi e restituzione virtuale di edifici protostorici multipiano, per esempio quelli elaborati a Creta nell'Età del Bronzo. In questi casi, pur essendo andato perduto l'elevato a causa dell'utilizzo di materiali deperibili, sussistono numerosi elementi documentari (strutture murarie conservate fino all'altezza dei primi piani, reperti in posizione di caduta dai piani superiori, ecc.) o indiziari (deduzioni sulla base di considerazioni architettoniche o funzionali) in grado di offrire informazioni sull'originaria conformazione dei piani superiori oggi perduti. La modellazione 3D, se praticata all'interno di una rigorosa cornice metodologica, rappresenta l'unico strumento in grado di gestire queste categorie di dati e offrire una nuova via per esplorare l'articolazione in elevato dell'architettura scomparsa in materiale deperibile.

PAROLE CHIAVE

3D modeling; virtual restoration; virtual archaeology; protohistoric multistorey architecture.

1. INTRODUZIONE

Nel variegato universo delle *digital humanities*, la modellazione 3D costituisce, da oltre un ventennio, uno dei campi favoriti di interazione tra nuovi strumenti digitali e archeologia, trovando le sue principali applicazioni nell'ambito della restituzione virtuale di architetture conservate solo parzialmente. Questo straordinario successo si spiega agevolmente visti gli innumerevoli vantaggi che esse offrono rispetto alle restituzioni grafiche tradizionali e, a maggior ragione, rispetto al restauro materico. Sintetizzando, tali vantaggi si possono ricondurre a tre punti principali: (1) creano modelli tridimensionali estremamente versatili per successive rielaborazioni grafiche; (2) hanno lo straordinario pregio della reversibilità rispetto alla restituzione materica, e (3) sono in grado di interagire con grande facilità con altri mezzi multimediali a finalità divulgative. Il loro utilizzo si inserisce coerentemente nella tradizione dei grandi modelli materici messi in opera con tecniche tradizionali, come il plastico di Roma Imperiale realizzato da Italo Gismondi nel 1933 per essere esposto al Museo della Civiltà Romana. Quest'ultimo ha costituito la base d'ispirazione per il progetto digitale di Bernard Frischer [3], *Rome Reborn*, finalizzato a proporre la medesima ricostruzione della monumentalità di Roma imperiale, stavolta però in formato digitale. A questa sono seguite altre imprese analoghe, varie per scala, area cronologica e geografica, dalla *Hierapolis Virtuale* [6] allo *Swedish Pompeii Project* [11].

L'impiego indiscriminato degli strumenti di modellazione per restituire contesti architettonici più o meno lacunosi comporta, però, rischi metodologici non indifferenti. Nella pratica dell'archeologia virtuale, le ricostruzioni tridimensionali, pur nell'ampio riconoscimento delle potenzialità delle loro applicazioni, solo raramente sono state pienamente integrate nella metodologia di ricerca [1: 43]; il loro uso è stato, piuttosto, applicato seguendo due principali linee di tendenza. La prima prende atto che le lacune della documentazione sono tali da non consentire ricostruzioni integralmente affidabili e si limita pertanto a proporre immagini tridimensionali verosimili ed accattivanti, ma senza alcuna pretesa di riprodurre con esattezza l'originale. Questo approccio è quello di gran lunga più praticato, soprattutto in ambito didattico e divulgativo, con evidenti applicazioni anche nel campo dell'industria turistica, ma tuttavia di scarso interesse scientifico. L'atteggiamento rientra in un filone di grande successo nell'ambito della modellazione 3D, talvolta incline al "pittoresco" e facilmente criticabile per la posizione rinunciataria e per l'informazione approssimativa e in certo grado fuorviante che trasmette ad un fruitore non specialista.

Una seconda linea di tendenza, più rigorosa e selettiva, si è diffusa soprattutto in ambito scientifico. Ispirata alle più avanzate riflessioni metodologiche sull'utilizzo della modellazione 3D nell'ambito della restituzione architettonica, punta

[1] Il contenuto del testo è il risultato di una elaborazione comune svolta da entrambi gli autori e dai medesimi condivisa in ogni sua parte; ad ogni modo, è presentato a firma distinta: Dario Puglisi ha scritto i paragrafi 1 ("Introduzione") e 3 ("Il progetto HTR 3D"), mentre Marco Chiricallo ha scritto i paragrafi 2 ("L'edilizia protostorica multipiano") e 4 ("Prospettive di sviluppo").

a ricreare immagini che abbiano un alto grado di affidabilità e verosimiglianza, sia in termini di volumi, che di cromatismo e finiture. Questa seconda tendenza viene di solito utilizzata come integrazione ed illustrazione del testo che descrive l'edificio antico, esplicitando i dati che consentono di ricostruirne con certezza l'aspetto originario. Essa è in linea con le più rigorose esigenze metodologiche, ma anche nelle sue applicazioni si possono riconoscere dei limiti. Il principale è che, proprio per rispondere al bisogno di essere quanto più attendibile possibile, essa finisce per essere applicabile solo ad edifici con lacune di limitata estensione, in modo che le parti da integrare siano ridotte al minimo e ipotizzabili con ampi margini di certezza. In secondo luogo, pur condotte in modo rigoroso e su basi in gran parte certe, anche queste ricostruzioni mantengono quel minimo grado di incertezza inevitabile in ogni restituzione di un edificio conservato in forma di rudere, con l'aggravante che l'aura di scientificità che le circonda ne offusca ulteriormente la natura pur sempre ipotetica.
In relazione al tema di questo convegno, va sottolineato come entrambe queste tendenze abbiano generalmente relegato la modellazione 3D a semplice procedura per la realizzazione di illustrazioni, più o meno attendibili, di una realtà scomparsa, demandando ad altra sede la riflessione critica, ovvero razionalmente e fattualmente fondata, sull'originale perduto [2; 9]. Quest'ultima avviene di solito a monte del processo di modellazione e viene esplicitata attraverso sistemi tradizionali, in genere resoconti scritti, mentre il modello viene elaborato solo al termine del procedimento ipotetico-deduttivo che ha portato a determinate conclusioni. Tale impostazione, che solo negli ultimi anni vede un tentativo di superamento, implica un netto ridimensionamento delle enormi potenzialità ermeneutiche che il "modello" elaborato in ambito digitale può acquisire nel momento in cui da semplice rappresentazione visiva tridimensionale, viene inteso come strumento di manipolazione di ipotesi al fine di acquisire avanzamenti di conoscenza. Secondo una dinamica quasi paradossale, questo aspetto metodologico risulta tanto più evidente quanto più esso è applicato a contesti in cui lo stato di conservazione dei resti archeologici non riesce a dare conto dell'elevato livello di complessità delle architetture originarie, soprattutto nel loro sviluppo in altezza. In tal senso, uno dei campi più favorevoli a dimostrare le potenzialità di questo approccio è costituito, dalle esperienze architettoniche multipiano in materiale deperibile sviluppatesi a Creta nel II millennio a.C.

2. L'EDILIZIA PROTOSTORICA MULTIPIANO

Con l'eccezione del fenomeno del megalitismo, l'architettura europea di età preistorica e protostorica fu realizzata facendo largo uso di materiali deperibili come il legno e l'argilla, i quali in contesto archeologico lasciano in genere poche tracce della loro presenza originaria. Il grado di complessità sociale ed avanzamento tecnologico implicò anche che le manifestazioni architettoniche di queste epoche più remote siano rimaste ad un livello relativamente semplice di articolazione spaziale, a maggior ragione rispetto all'elevato degli edifici. Una eccezione a questo quadro è rappresentata dalle esperienze architettoniche fiorite a Creta nel corso del II millennio a.C. [8]. La cosiddetta "civiltà minoica palaziale" fu infatti caratterizzata da un livello tecnologico estremamente avanzato se paragonato al resto d'Europa, che si manifestò con particolare evidenza, come a tutti è noto, nell'architettura monumentale. Meno noto è, invece, che almeno a partire dal III millennio a.C., gli edifici costruiti dai Minoici, non solo a carattere monumentale ma anche domestico, erano sempre caratterizzati da una articolazione su più piani collegati da scale. La terribile eruzione avvenuta a Thera alla metà del II millennio a.C., grazie alle eccezionali modalità di distruzione, ne ha conservato testimonianza eclatante, con strutture superstiti fino a tre livelli sovrapposti [7]. Nel resto di Creta, tuttavia, e in particolare negli edifici monumentali tradizionalmente noti come "palazzi", questa porzione così rilevante del costruito è quasi del tutto scomparsa a causa del fatto che era in gran parte realizzata in legno, argilla e mattone crudo. Pochi resti riferibili a questi livelli più alti sopravvivono, sotto forma di parti di murature conservate oltre il livello del solaio, o come oggetti le cui quote di rinvenimento consentono di ipotizzarne la giacitura originaria ai piani superiori. Questo stato della documentazione ha fatto sì che gli archeologi, a parte qualche coraggioso tentativo, abbiano sostanzialmente rinunciato a conoscere queste porzioni degli edifici minoici, ritenendo impossibile acquisire informazioni attendibili su di essi. Le limitazioni che un tale atteggiamento pone ad una piena comprensione delle esperienze architettoniche fiorite a Creta nel II millennio sono evidenti. Meno esplicite sono le potenzialità che un uso rigoroso e metodologicamente strutturato della modellazione 3D può avere per superare questa impasse.

3. IL PROGETTO HTR 3D

Una sperimentazione sulle architetture minoiche di età neopalaziale conservate nel sito di Haghia Triada e risalenti al Tardo Minoico IB (1550-1490 a.C.) è stata condotta a partire dal 2018 dall'Università di Catania e dal Politecnico di Bari in accordo con la Scuola Archeologica Italiana di Atene [10].

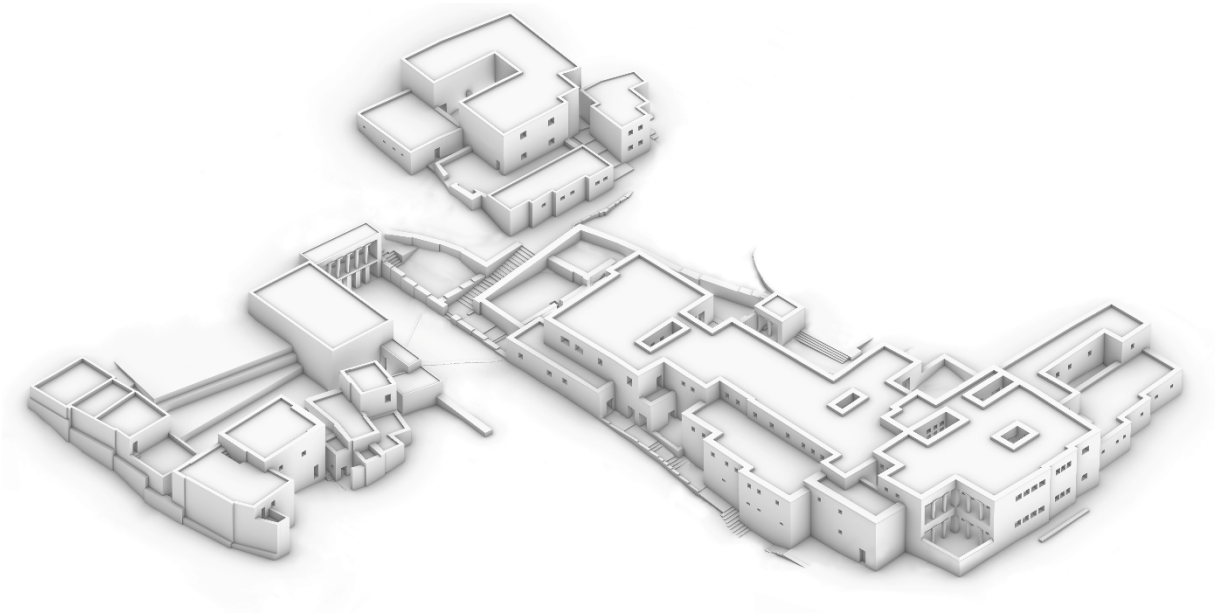


Figura 1. Prima versione della ricostruzione tridimensionale dell'insediamento Tardo Minoico IB di Haghia Triada, realizzata con la collaborazione di A. Cascione, D. Chiricallo, A. Labbattaglia, I. Leone, M. T. Lence, V. Liuzzi, M. Corona e M. Delfino

Denominata "Progetto HTR 3D", tale sperimentazione ha provato a superare la posizione rinunciataria invalsa nell'ambito degli studi di settore riguardo alla ricostruzione dei piani superiori degli edifici minoici, ricorrendo all'uso sistematico della modellazione 3D, intesa, come già detto, non come semplice tecnica di rappresentazione grafica, ma come strumento dinamico in grado di facilitare l'elaborazione e verifica di ipotesi ricostruttive.

Il contesto archeologico scelto per la sperimentazione si presta ad un tale approccio per una serie di ragioni: innanzitutto, si tratta di un complesso assegnabile ad un arco di tempo relativamente ristretto; in secondo luogo, i resti presentano un'estensione ed uno stato di conservazione eccellenti se confrontati con gli standard coevi. In terzo luogo, la disposizione delle strutture su più terrazze degradanti consente di assicurarne l'articolazione su almeno due piani, di cui il superiore delle strutture a valle si pone orientativamente allo stesso livello del pianterreno a monte. Infine, l'intero complesso, portato in luce all'inizio del secolo scorso e indagato ulteriormente tra il 1977 e il 2012 sotto la direzione di V. La Rosa, è già stato in gran parte pubblicato con un notevole grado di dettaglio [5, 9].

Il progetto si è servito di software comuni e di facile reperibilità: Rhinoceros per la modellazione 3D, su una base planimetrica in CAD delle strutture, ed ha riguardato l'intero insediamento, costituito dalla cosiddetta "Villa Reale", grande edificio monumentale esteso circa 1500 m^2, circondata dal cosiddetto "Villaggio", modesto agglomerato di poche unità domestiche esteso circa 1100 m^2 (vd. Fig. 1). La fase preliminare del progetto ha richiesto un poderoso sforzo di rielaborazione dei dati già acquisiti, che è stato talvolta necessario integrare con nuovi rilievi sul campo. In questa fase, l'obiettivo principale è stato quello di ricostruire la struttura multipiano scomparsa dei vari edifici coinvolti, individuando precisi indicatori archeologici della presenza dei piani superiori e formulando principi di massima su cui basare la restituzione. In particolare, l'originaria presenza di piani superiori è ipotizzata sulla base dei seguenti indicatori archeologici: porzioni di strutture superstiti riferibili ai primi piani; presenza di scale; elementi architettonici del piano terra finalizzati al sostegno dei piani superiori; rinvenimento di materiali costruttivi e oggetti crollati dal primo piano; articolazione a terrazze e dialogo fra piani di calpestio interni ed esterni; considerazioni di tipo statico-costruttivo.

A conclusione della fase preliminare si è anche posto il problema della resa grafica del modello. Si è optato per una rappresentazione minimalista, che puntasse ad evidenziare la dimensione volumetrica delle architetture restituite, senza aggiungere effetti realistici che avrebbero ulteriormente aumentato la componente ipotetica della rappresentazione. L'intento della versione prescelta è stato quello di trametterе una immagine immediata e leggibile della tridimensionalità dell'edificio, in grado di renderne intellegibili gli aspetti morfologici e funzionali, esprimendo al tempo stesso la sua natura ipotetica attraverso la resa non realistica.

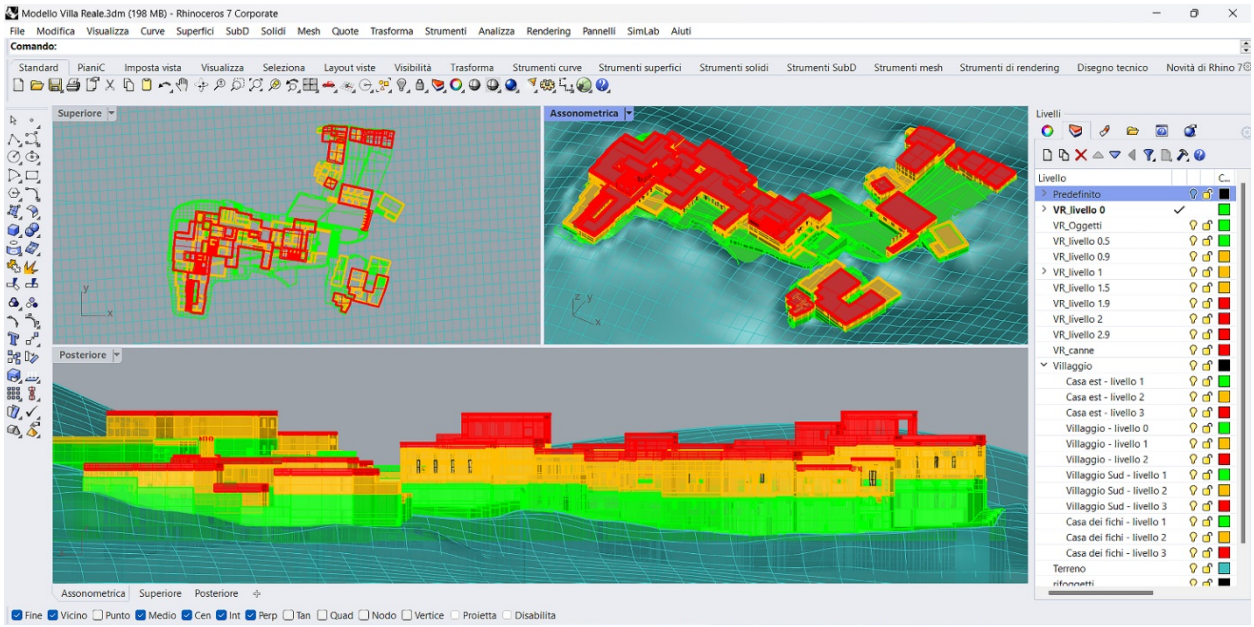


Figura 2. Schermata del programma Rhinoceros utilizzato per realizzare il secondo modello tridimensionale

Una volta modellizzata questa prima massa di dati e le relative ipotesi di restituzione, ci si è concentrati sulle architetture della Villa, sottoponendo a verifica sistematica le ricostruzioni proposte nella fase preliminare. In questa fase avanzata, la restituzione di ciascun vano e settore dell'edificio è stata condotta utilizzando il modello come un vero e proprio sviluppatore di ipotesi, definendo quindi ciascuna di quelle possibili e selezionando, attraverso serrate argomentazioni, quelle più plausibili dalle incerte e dalle inattendibili. Si è giunti così alla elaborazione di un nuovo modello di restituzione della Villa, piuttosto diverso da quello proposto nella prima fase, che compendia in sé tutte le soluzioni ricostruttive ritenute più probabili, suddiviso in layer corrispondenti ai differenti piani in modo che fosse più facilmente esplorabile (vd. Fig. 2). Anche in questa fase si è scelto di mantenere la resa grafica elaborata per il modello precedente (vd. Fig. 3).
Nonostante la gran parte delle ricostruzioni proposte per i piani superiori abbia una base ipotetica più o meno ampia, vi è un indubbio zoccolo duro di acquisizioni certe o ampiamente probabili circa l'estensione complessiva dei piani superiori, le quote di collocazione dei medesimi, i sistemi di comunicazione interni ed esterni, le relazioni progettuali e strutturali tra piano terra e primo piano, che costituiscono un poderoso passo avanti nella conoscenza dell'edificio originario inteso nella sua interezza e non nella sola porzione relativa al piano inferiore. È inutile sottolineare le potenzialità intrinseche in questa procedura, qualora essa sia applicata agli innumerevoli altri edifici minoici multipiano, monumentali e non, finora studiati solo a livello del piano terra. Tuttavia, a questo punto della ricerca, sono emersi ulteriori limiti del lavoro già svolto, che aprono altrettante prospettive per sviluppi futuri che sarà opportuno brevemente tratteggiare in questa sede.

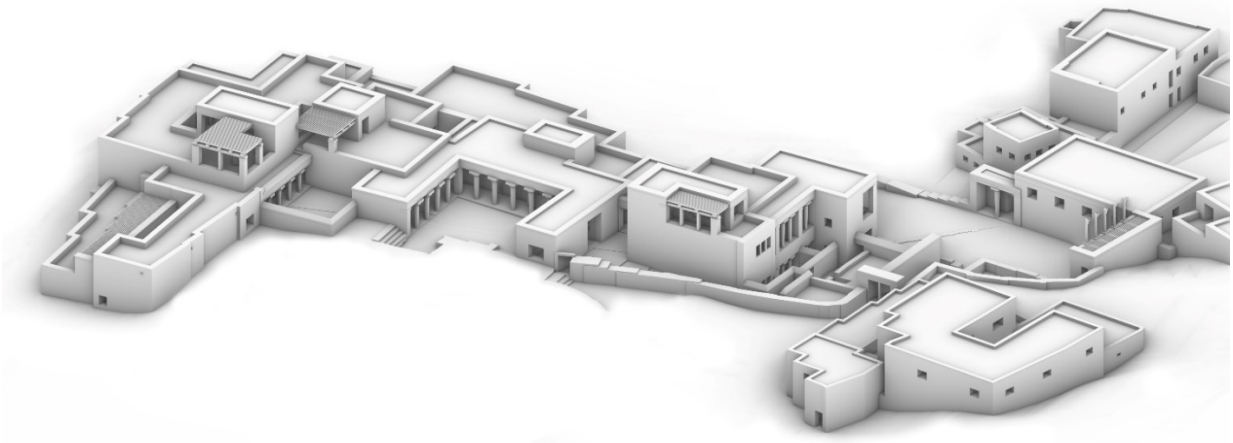


Figura 3. Seconda versione della ricostruzione tridimensionale della Villa Reale di Haghia Triada nel Tardo Minoico IB

4. PROSPETTIVE DI SVILUPPO

Nonostante l'applicazione di un approccio tridimensionale allo studio dell'architettura minoica abbia dimostrato con il caso studio di Haghia Triada tutte le sue potenzialità scientifiche, l'impiego di una modellazione 3D di tipo tradizionale ha evidenziato inevitabilmente anche i suoi limiti. Allo stato attuale, infatti, i modelli finora realizzati possono essere utilizzati e trasmessi al pubblico solo sotto forma di immagini bidimensionali (o tridimensionali nel caso del modellino della Villa Reale presentato alla mostra "*Nell'isola di Dedal*o", Catania 22 maggio - 30 giugno 2023) illustrative di testi scritti che presentano i dati, discutono le argomentazioni a sostegno delle diverse ipotesi e selezionano quelle ritenute più plausibili. In questi contesti, il modello 3D perde però la sua natura dinamica di strumento per la fruizione dello spazio tridimensionale scomparso e per la elaborazione e verifica di ipotesi circa la sua originaria morfologia. A questo primo limite se ne affianca un secondo, ovvero il fatto che il modello non è in grado di conservare e trasmettere l'insieme dei dati e delle affermazioni ipotetico-deduttive sulla base delle quali è stato costruito. Tutta questa massa di dati, infatti, è al momento conservata su formati differenti e non è fruibile da chi utilizza il modello.

I limiti della modellazione tradizionale come strumento di rappresentazione, anche quando integrata nel processo ricostruttivo, sono in effetti ormai ben noti alla comunità scientifica [1]. In questo senso, iniziative quali la Carta di Londra[2] e i Principi di Siviglia[3] hanno tentato di impostare una cornice metodologica condivisa, all'interno della quale si sono sviluppate soluzioni differenziate sia nella metodologia sia negli strumenti informatici impiegati; ancora lontane da una standardizzazione, tali soluzioni si evolvono con la rapidità e il fermento dei diversi strumenti informatici su cui si basano. Fra le varie strade percorribili, alcune destano particolare interesse per compiere un ulteriore passo avanti anche rispetto allo studio tridimensionale dell'architettura minoica.

In particolare, spunti interessanti per superare il limite legato alla condivisibilità e verificabilità delle ipotesi proposte potrebbero venire dal recente progetto "*Rome Transformed* - SCIEDOC"[4]. In questo caso, le restituzioni tridimensionali, non a caso definite "provocazioni", sono proposte non come un esito finale, ma come uno stato intermedio consapevolmente provvisorio e potenzialmente inesatto di un processo di studio dei dati archeologici che, adeguatamente documentato, può provocare, appunto, il dibattito critico fra gli studiosi. Ciascuna proposta ricostruttiva è pubblicata online corredata dai dati e dai processi interpretativi che l'hanno giustificata, anche in più versioni diverse. Per quanto questo approccio costituisca un notevole passo avanti verso l'applicazione dei principi FAIR (*findable, accessible, interoperable, reusable*), i modelli realizzati continuano ad essere però presentati solo nella loro forma finale, senza la possibilità di una reale interattività e di una loro interrogabilità diretta. Per superare questo secondo limite occorre che siano rese accessibili non le riproduzioni bidimensionali del modello, ma il modello stesso, ed inoltre che quest'ultimo sia strutturato come un vero e proprio database che gestisce i dati su cui si basa la restituzione proposta.

La combinazione tra modellazione 3D e funzioni database ha già conosciuto diverse interessanti sperimentazioni, tra le quali ne vanno segnalate almeno due per le evidenti applicazioni che possono avere allo studio dell'architettura protostorica multipiano. La prima è costituta dall'approccio dell'*Extended Matrix* [1]: fortemente ancorata al metodo stratigrafico, si tratta di una "estensione" del *matrix* di Harris, che include oltre alle unità stratigrafiche tradizionali anche unità stratigrafiche virtuali, corrispondenti ad elementi tridimensionali ricostruiti come reintegrazioni di altri sopravvissuti (USV/*Structural*) oppure ipotizzabili sulla base di fonti esterne (USV/*Non structural*). Ciascun elemento del grafico è connesso ad un ambiente tridimensionale, esplicitando fonti e processo ricostruttivo, e rendendo possibile sia la creazione di modelli semplificati e interrogabili, anche con gradienti differenziati delle varie unità in base al loro grado di affidabilità [2; 5], sia, in seconda battuta, di modelli fotorealistici, impiegabili anche per le visite virtuali. Di interesse sono anche le potenzialità dell'implementazione nel processo di lavoro delle tecniche BIM (*Building Information Modeling*) applicate ai beni culturali (HBIM) e, in particolare, ai contesti archeologici [4]. Tali applicazioni, in fase ancora sperimentale sebbene in grande crescita, hanno già dimostrato la loro utilità nell'associare alle varie parti del modello tridimensionale attributi qualitativi, numerici e di altra natura, consentendo la comparazione di dati eterogenei e una maggiore facilità di analisi sul modello. Rispetto alle specifiche esigenze della ricostruzione archeologica, questa seconda opzione sconta l'assenza di una articolazione per US che è invece presente nel sistema *Extended Matrix* e che si adatta più facilmente alle procedure stratigrafiche tipiche dell'approccio archeologico. Di contro, però, consente di associare più agevolmente analisi sui comportamenti strutturali delle parti già conservate a verifiche della sostenibilità delle ipotesi di restituzione. In prospettiva, l'utilizzo combinato dei due sistemi costituirebbe un poderoso strumento di avanzamento nello studio dell'architettura multipiano protostorica cretese.

[2] https://londoncharter.org/
[3] http://sevilleprinciples.com/
[4] https://research.ncl.ac.uk/rometrans/

In conclusione, il progetto HTR 3D rappresenta una prova evidente di come la modellazione 3D possa aprire nuove inedite prospettive di conoscenza anche se applicata a contesti documentari, come il sito di Haghia Triada, già noti e oggetto di studio da oltre un secolo, e conferma le ulteriori potenzialità che tale approccio può assumere se combinato con altre metodologie di gestione e analisi dei dati in ambiente digitale.

5. RINGRAZIAMENTI

La parte preliminare della sperimentazione è stata svolta nell'ambito del progetto "STORAGE. Dai dati al Web" (Programma Pia.ce.ri), dell'Università di Catania negli anni 2020-22. Il lavoro, a partire dal 2023, è in corso di sviluppo nell'ambito del progetto CHANGES, *Spoke* 6, *History, Conservation, Restoration of Cultural Eritage*.

BIBLIOGRAFIA

[1] Demetrescu, Emanuel. «Archaeological stratigraphy as a formal language for virtual reconstruction. Theory and practice». *Journal of Archaeological Science* 57 (2015): 42–55.

[2] Demetrescu, Emanuel, e Daniele Ferdani. «From Field Archaeology to Virtual Reconstruction: A Five Steps Method Using the Extended Matrix». *Appl. Sci* 11, fasc. 5206 (2021): 1–23.

[3] Frischer, Bernard. «Cultural and Digital Memory: Case studies from the Virtual World Heritage Laboratory». *Memoria Romana: Memory in Rome e Rome in Memory* 10 (2014): 151–62.

[4] Gaiani, Marco, Simone Garagnani, Andrea Gaucci, e Paola Moscati. *ArchaeoBIM. Theory, Processes and Digital Methodologies for the Lost Heritage*. Bologna: BONONIA UNIVERSITY PRESS, 2021.

[5] Halbherr, Federico, Enrico Stefani, e Lucia Banti. «Haghia Triada nel periodo tardo-palaziale». *Annuario della Scuola Archeologica Italiana di Atene* 55 (1977): 7–296.

[6] Limoncelli, Massimo. «Archeologia Virtuale a Hierapolis di Frigia: la restituzione dell'immagine della città». In *Theatroeideis. L'immagine della città, la città delle immagini*, II:217–34. Thiasos Monografie, 2018.

[7] Palyvou, Clairy. *Akrotiri, Thera: An Architecture of Affluence 3,500 Years Old*. Philadelphia: Institute for Aegean Prehistory, 2005.

[8] Palyvou, Clairy. *Daidalos at work. A phenomenological approach to the study of Minoan architecture*. Philadelphia: INSTAP Academic Press, 2018.

[9] Puglisi, Dario. «Haghia Triada nel periodo Tardo Minoico I». *Creta Antica* 4 (2003): 145–98.

[10] Puglisi, Dario, e Marco Chiricallo. «Modellazione 3D ad Haghia Triada: note per un approccio tridimensionale all'architettura minoica». *Thiasos* 12 (2023): 309–25.

[11] Touati, Anne-Marie Leander, Thomas Staub, e Renée Forsell. «From 2D and 3D documentation to 4D interpretation». *OpAthRom* 14 (2021): 181–226.

MLS con sensore LiDAR *Apple* per lo scavo archeologico: applicazioni pratiche

Luigi M. Caliò[1], Antonello Fino[2], Gian Michele Gerogiannis[3]
[1] DISUM, Università di Catania, Italia - luigi.m.calio@gmail.com
[2] ArCoD, Politecnico di Bari, Italia - antonello.fino@poliba.it
[3] DISUM, Università di Catania, Italia - g.gerogiannis@unict.it

ABSTRACT
Nel poster si esamina l'impatto delle nuove tecnologie, in particolare il sensore LiDAR di *Apple*, sull'archeologia, evidenziando l'evoluzione nella documentazione grafica di scavo. L'analisi autoptica delle testimonianze archeologiche, integrata con metodologie avanzate, costituisce un processo fondamentale per la comprensione e la ricostruzione dei contesti. Si intende dimostrare come l'integrazione del LiDAR e di tecniche *image-based* abbia ottimizzato la raccolta dati, pur sottolineando le limitazioni in termini di precisione rispetto ai metodi tradizionali. Nonostante ciò, l'efficacia del LiDAR nel documentare in modo rapido ed economico le fasi di scavo e i reperti archeologici rappresenta un significativo progresso metodologico. Si discute dell'esperienza diretta nelle campagne di scavo e dell'adozione di questa tecnologia in contesti diversi, puntando alla creazione di una documentazione archeologica più accessibile e dettagliata, evidenziando i vantaggi in termini di velocità e costi, nonché le potenzialità divulgative dei modelli virtuali generati.

PAROLE CHIAVE
Archeologia digitale; rilievo archeologico; sensore LiDAR Apple, MLS (Mobile Laser Scanner).

1. INTRODUZIONE
Come noto, l'attività di documentazione grafica in ambito archeologico non può prescindere da un primo indispensabile passaggio che trova il suo compimento nell'analisi autoptica della materia antica [3: 75]. Questo processo, incentrato sul riconoscimento e la comprensione delle testimonianze di azioni antropiche, rappresenta il punto di partenza inevitabile per l'interpretazione e le successive ipotesi di ricostruzione dei vari contesti analizzati, nelle loro fasi differenti di vita. Le informazioni ottenute durante un'indagine, risultato di osservazioni meticolose e informate, dipendono essenzialmente da una solida formazione teorico-pratica della disciplina archeologica, e si avvalgono da sempre degli strumenti del rilievo come supporto fondamentale, definendo e guidando le tappe dell'acquisizione di dati metrologici e morfologici nonché della loro restituzione grafica. Di conseguenza, l'attività di documentazione grafica si configura come un momento centrale per l'acquisizione della conoscenza e si presenta come fase critica che richiede massima precisione, stabilendo le basi per il successivo processo di indagine [3]. È importante sottolineare che, in questo senso, lo sviluppo tecnologico degli ultimi decenni ha significativamente influenzato le modalità operative, specialmente nel contesto archeologico, agevolando notevolmente le operazioni di registrazione e documentazione, attività che rappresentano un segmento significativo, in termini temporali ed economici, del lavoro sul campo. Nonostante ciò, la conoscenza e l'applicazione delle tecniche di rilievo tradizionali rimangono cruciali per qualsiasi ricerca in ambito archeologico, poiché la capacità di produrre, decifrare e interpretare piante e sezioni nelle adeguate scale di rappresentazione costituisce un'abilità insostituibile, che non può essere completamente affidata nemmeno agli strumenti di misurazione più avanzati e precisi attualmente disponibili.
È ormai diffuso e ampiamente accettato in questo senso l'uso delle tecniche *imaged based*, per la produzione di immagini fotogrammetriche di piante, sezioni e prospetti, ma anche di reperti ceramici o frammenti architettonici, che ha consentito di portare a compimento programmi di lavoro altrimenti destinati a più lunghe permanenze, non sempre compatibili con le attuali organizzazioni e mezzi economici degli enti universitari e di ricerca [7, 9, 12]. Inoltre, la facilità di reperimento degli strumenti di acquisizione, quali fotocamere digitali e accessori come aste telescopiche e sistemi di scatto in remoto, nonché dei sempre più accessibili sistemi UAS, nonché dei *software* per la post-elaborazione in *camera-scanner*, molti dei quali *open source* [5], hanno consentito un'ampia diffusione della metodologia che, di fatto, ha notevolmente ridimensionato la concorrenza degli scanner laser, strumentazioni più onerose sia in termini economici, sia per quanto riguarda l'archiviazione dei dati elaborati, fermo restando applicazioni specifiche sia di questi ultimi, sia delle insostituibili stazioni totali e sistemi GNSS [1]. Metodi e tecnologie diverse, quindi, da scegliere in base alle necessità della ricerca, con l'obiettivo di preservare la possibilità che i dati e i risultati ottenuti possano essere a mano a mano integrati, con il fine di avere una documentazione completa, accessibile e il più precisa possibile. In questo senso, è opportuno ribadire, però, che le fasi del rilievo non sono una mera elaborazione di dati da affidare a un tecnico, bensì si tratta di analisi che devono essere condotte da operatori specializzati e pienamente consapevoli della natura delle emergenze [2]. Si tratta di un presupposto

Figura 1. Agrigento, scavo del teatro. Comparazione fra l'elaborazione LiDAR (in alto), fotogrammetria imaged-based (centro) e rilievo di dettaglio tradizionale (Politecnico di Bari).

fondamentale, ad esempio, nelle fasi di documentazione di uno scavo, che avvengono contestualmente alle operazioni di asportazione di terra e alla messa in luce di nuove stratigrafie, sistematicamente documentate.

Questo contributo espone, dunque, esperienze dirette di campagne di scavo, svolte fino al 2022 all'interno del progetto di ricerca "STORAGE. Dai dati al Web" (Programma Pia.ce.ri) e dal 2023 nell'ambito del progetto CHANGES, Spoke 6, in contesti in cui si sono resi necessari lavori di pulizia e preparazione dell'area da indagare e a volte, a fine scavo, anche operazioni di colmatura e chiusura dei saggi aperti. Più nel dettaglio, si intende sottolineare come la ricerca di sistemi in grado di redigere una documentazione grafica in tempi rapidi, ma che salvaguardassero l'attendibilità del dato, sia stata essenziale al fine di ottenere risultati scientificamente validi. Si è cercato altresì di individuare i metodi più idonei a documentare graficamente un'area di scavo archeologico, tenendo come parametri base la precisione, l'efficacia e la velocità di esecuzione dell'elaborato, senza trascurare l'aspetto economico. È in questo senso che nelle recenti missioni archeologiche condotte dal DISUM dell'Università degli Studi di Catania e dall'ArCoD del Politecnico di Bari ad Agrigento, a Ostia Antica, ma anche a Pompei, è stato sperimentato l'utilizzo di scanner portatili per il rilevamento dinamico terrestre MLS (*Mobile Laser Scanner*) che utilizzano il sensore LiDAR, che *Apple* ha installato sui suoi dispositivi mobili a partire dal 2020.

2. LiDAR APPLE E APPLICAZIONE SUL CAMPO

Sebbene sia stato concepito per applicazioni relative al settore della realtà aumentata, a partire dalla sua comparsa sul mercato, diverse aziende hanno sviluppato app che ne sfruttano le potenzialità di misurazione per generare modelli tridimensionali, da cui è poi possibile estrapolare immagini *raster* dell'oggetto rilevato. Il sensore, utilizzabile in modalità *indoor* e *outdoor*, ha una gittata di m 4.90 e funziona in associazione alla fotocamera che attribuisce ai punti della nuvola acquisita informazioni RGB circa la loro colorazione. Sin dall'inizio, lo strumento ha destato l'interesse di addetti ai lavori nel settore della documentazione grafica, con applicazioni in diversi settori, fra cui la geologia e l'architettura, ma allo stesso tempo, le potenzialità del sensore sono state oggetto di curiosità anche fra chi opera in campo archeologico. Allo stesso modo, negli ultimi anni diverse pubblicazioni hanno esposto l'esito di una serie di sperimentazioni, con i relativi risultati in termini di vantaggi e di errore [4, 6, 8, 10, 11, 13, 14]. Partendo dall'analisi della bibliografia sul tema e grazie alla possibilità di usufruire della tecnologia utilizzando un iPad Pro di quarta generazione, l'équipe di ricerca ha potuto verificare sul campo alcune applicazioni. L'acquisizione dei dati è avvenuta attraverso l'utilizzo delle *features* gratuite dell'app Polycam, che genera una superficie *mesh* a rete di poligoni

texturizzata misurabile già all'interno dell'app ed esportabile in formato OBJ per ulteriori elaborazioni con i principali software di modellazione.
Primi test sono stati effettuati con l'intento di verificare l'attendibilità del modello 3D rispetto a quello prodotto con tecnica *imaged based* tradizionale, i cui tempi di elaborazione, sebbene l'alto livello di precisione, sono significativamente elevati. Nel caso del mosaico rinvenuto nel vano absidato presso il settore settentrionale del teatro di Agrigento, il modello 3D acquisito con sensore LiDAR, scansionando ad una distanza mai superiore ai m 1.5, ha prodotto un'immagine piuttosto fedele dello spazio, che ha consentito di apprezzare anche numericamente e qualitativamente le singole tessere musive, restituendo però un errore compreso fra i cm 2 e 3 in pianta, rispetto alle misurazioni effettuate con stazione totale (vd. Fig. 1). Lo stesso rilievo effettuato con fotogrammetria tradizionale ha restituito errori sub millimetrici, portando a concludere che per rilievi di dettaglio planimetrici in cui è richiesto un elevato fattore di precisione, l'utilizzo del LiDAR *Apple*, non può essere la prima scelta fra le varie tecniche di acquisizione. Questa esperienza ha però dato modo di comprendere che con le stesse modalità di acquisizione, un errore di quel tipo è senz'altro accettabile nel caso della documentazione stratigrafica in fase di scavo. Nel caso dello scavo di Ostia (campagna 2023), pertanto, si è ritenuto opportuno effettuare piante e sezioni stratigrafiche, che considerati i tempi di elaborazione, hanno senza dubbio agevolato il lavoro sul campo, restituendo un'immagine dettagliata dello stato dei luoghi. La facilità di operazione, inoltre, ha comportato un significativo incremento della documentazione delle operazioni, che ha certamente aiutato il lavoro degli archeologi che in più casi sono costretti ad operare delle scelte irreversibili, ma che grazie a questo sistema, possono di fatto "congelare" alcuni passaggi cruciali per le attività di scavo. Non solo, la possibilità di estrapolare immagini *raster* dal modello 3D, è fondamentale nella redazione dei diari di scavo e l'utilizzo stesso del tablet, quale strumento di scrittura e annotazione, semplifica il non facile rapporto con lo schizzo a mano libera che in alcuni casi può generarsi fra i responsabili di saggio, uniformando la documentazione con un più oggettivo strumento di rappresentazione (vd. Fig. 2).

Figura 2. Esempio di applicazione del LiDAR per la documentazione di scavo presso il Capitolium di Pompei (elaborazione Politecnico di Bari)

Un utilizzo della tecnologia ancora da sviluppare e che restituisce esiti non particolarmente felici, riguarda la scansione di oggetti di piccole dimensioni o particolarmente dettagliati come reperti ceramici, frammenti architettonici, monete, ecc.
A corollario delle attività dirette sul campo vi è, infine, un'ulteriore finalità riguardo l'applicazione di questa metodologia che interessa l'aspetto divulgativo delle ricerche in corso. Grazie, infatti, alla realizzazione di modelli virtuali [2] delle aree

di scavo è possibile mostrare aspetti dello scavo non sempre accessibili ai visitatori, soprattutto per motivi di sicurezza (vd. Fig. 3).

Figura 3. Esempio di modellazione 3D. Strutture presso lo scavo di Ostia antica (elaborazione Politecnico di Bari).

Spesso, difatti, la messa in sicurezza del cantiere di scavo non prevede la fruizione da parte dei visitatori, anche dopo che le operazioni di scavo sono terminate. I modelli virtuali, quindi, possono svolgere un ruolo significativo nel promuovere e diffondere la conoscenza di un sito archeologico in fase di scavo, fornendo rappresentazioni visive che rendono più accessibili le scoperte archeologiche, altrimenti oscure.
In questi termini il presente poster affronta a più livelli le tematiche discusse nella conferenza. Da un lato, infatti, la creazione di questi modelli virtuali permette di preservare la memoria collettiva documentando ogni fase del processo di scavo, che, come già ribadito, costituisce sempre un'azione di per sé distruttiva; dall'altro, facilita la creazione di uno spazio virtuale di condivisione del lavoro archeologico e delle sue scoperte, inclusi elementi come strutture, stratigrafie, artefatti, che emergono progressivamente durante l'indagine.
Infine, il bilancio sull'utilizzo di questa nuova strumentazione è sostanzialmente positivo. Al netto della scarsa precisione a scale di rappresentazione di dettaglio, l'uso che se ne può fare nel corso di una campagna di scavo, lo rende uno strumento che presto si confermerà ampiamente condiviso nel settore, cosa che porterà ad ulteriori sviluppi della tecnologia, ancora giovane e con ampi margini di miglioramento, specifici per l'archeologia.

3. RINGRAZIAMENTI

La parte preliminare della sperimentazione è stata svolta nell'ambito del progetto "STORAGE. Dai dati al Web" (Programma Pia.ce.ri), dell'Università di Catania negli anni 2020-22. Il lavoro, a partire dal 2023, è in corso di sviluppo nell'ambito del progetto CHANGES, Spoke 6, History, Conservation, Restoration of Cultural Eritage.

BIBLIOGRAFIA

[1] Alvaro, Corrado, Valentina Albano, Simone Amici, Jade Bajeot, Valeria Danesi, Gian Michele Gerogiannis, Chiara La Marca, et al. «The shape of monuments project. Current activities and technological training in university – industry partnership». *Scienze dell'Antichità* 22 (2016): 213–34.

[2] Barceló, Juan A. «Virtual Reality for Archaeological Explanation. Beyond "Picturesque" Reconstruction». *Archeologia e Calcolatori* 12 (2001): 221–44.

[3] Bianchini, Marco. *Manuale di rilievo e di documentazione digitale in archeologia*. 1a edizione. Scienze dell'antichità, filologico-letterarie e storico-artistiche; 362. Roma: Aracne, 2008.

[4] Cohen-Smith, Hannah, Simon H. Bickler, Benjamin Jones, Bernie Larsen, e Aaron Apfel. «New Tech for Old Jobs: Handheld LiDAR for Feature Recording». *Archaeology in New Zealand* 65, fasc. 2 (2022): 14–27.

[5] Dabove, Paolo, Nives Grasso, e Marco Piras. «Smartphone-Based Photogrammetry for the 3D Modeling of a Geomorphological Structure». *Applied Sciences* 9, fasc. 18 (2019): 3884. https://doi.org/10.3390/app9183884.

[6] Donlic, Matea, Tomislav Petkovic, e Tomislav Pribanic. «On Tablet 3D Structured Light Reconstruction and Registration». In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2462–71, 2017. https://doi.org/10.1109/ICCVW.2017.290.

[7] Fino, Antonello. «Il rilievo del circuito murario di Byllis». In *Byllis. La città e il territorio. Ricerche 2014-2017. II. La documentazione grafica*, (a cura di) Roberta Belli Pasqua e Luigi M. Caliò, 17–22. Thiasos, Monografie 14. Roma: Edizioni Quasar, 2018.

[8] Fiorini, Andrea. «Scansioni dinamiche in archeologia dell'architettura: test e valutazioni metriche del sensore LiDAR di Apple». *Archeologia e Calcolatori* 33, fasc. 1 (2022): 35–54.

[9] Gilboa, Ayelet, Ayellet Tal, Ilan Shimshoni, e Michael Kolomenkin. «Computer-based, automatic recording and illustration of complex archaeological artifacts». *Journal of Archaeological Science* 40, fasc. 2 (febbraio 2013): 1329–39. https://doi.org/10.1016/j.jas.2012.09.018.

[10] Günen, Mehmet Akif, İlker Erkan, Şener Aliyazıcıoğlu, e Cavit Kumaş. «Investigation of geometric object and indoor mapping capacity of Apple iPhone 12 Pro LiDAR». *Mersin Photogrammetry Journal* 5, fasc. 2 (2023): 82–89. https://doi.org/10.53093/mephoj.1354998.

[11] Luetzenburg, Gregor, Aart Kroon, e Anders Anker Bjørk. «Evaluation of the Apple iPhone 12 Pro LiDAR for an Application in Geosciences». *Scientific Reports* 11 (2021). https://doi.org/10.1038/s41598-021-01763-9.

[12] Remondino, Fabio, e Stefano Campana, (a cura di). *3D recording and modelling in archaeology and cultural heritage: theory and best practices*. International Series 2598. Oxford: Archaeopress BAR, 2014.

[13] Spreafico, Alessandra, Filiberto Chiabrando, Giulio Fabio Tonolo, e Lorenzo Teppati Losè. «Apple iPad Pro: test e valutazioni metriche sul sensore LiDAR integrato». In *#AsitaAccademy2021*, 421–23, 2009.

[14] Teppati Losè, Lorenzo, Alessandra Spreafico, Filiberto Chiabrando, e Fabio Giulio Tonolo. «Apple LiDAR Sensor for 3D Surveying: Tests and Results in the Cultural Heritage Domain». *Remote Sensing* 14, fasc. 17 (2022). https://doi.org/10.3390/rs14174157.

Odonimi d'Italia e Digital Public History: le problematiche di una schedatura partecipata

Enrica Salvatori[1], Vittore Casarosa[2], Riccardo Chiari[3]
[1] Università di Pisa, Italia - enrica.salvatori@unipi.it
[2] CNR Istituto di Scienza e Tecnologie dell'Informazione, Italia - casarosa@isti.cnr.it
[3] Università di Pisa, Italia - r.chiari@studenti.unipi.it

ABSTRACT
Si presentano le impostazioni metodologiche di un progetto sulla schedatura partecipata degli odonimi d'Italia al fine di avviare un utile confronto con la comunità scientifica degli umanisti digitali. Per quanto ancora in una fase iniziale di elaborazione e sperimentato per una sola città italiana, riteniamo importante spiegare gli intenti, i metodi e le problematiche che si incontrano nella categorizzazione degli odonimi del territorio nazionale ed evidenziare le difficoltà e opportunità nella scelta di lavorare su un livello locale, anche e soprattutto in previsione del ricorso ineludibile alla partecipazione diretta del pubblico in attività di *Citizen Humanities*.

PAROLE CHIAVE
Odonomastica; toponomastica; Wikidata; Digital Public History.

1. INTRODUZIONE
A differenza dei toponimi, gli odonimi sono frutto di atti politici, che hanno avuto nella storia della nostra nazione alcune fasi di creazione e cancellazione particolarmente rilevanti: l'unità d'Italia, il regime fascista, l'istituzione della Repubblica. Anche se la pratica di intitolare strade, piazze e altri luoghi/istituti pubblici è presente negli stati preunitari, è comunque con l'unità d'Italia che si ha la prima profonda rielaborazione del processo risorgimentale con una sua "scrittura" nel tessuto urbano, unita alla creazione di un calendario civile e parallelamente di un *pantheon* di modelli umani atti a correttamente ricordare e rappresentare la storia patria [2, 6]. La gestione di questo patrimonio, a un tempo materiale e immateriale, è diventata sempre più importante a livello eminentemente cittadino, in quanto vi si incrociano processi diversi: da un lato il ruolo dell'amministrazione comunale che stabilisce le strutture da nominare e decide ufficialmente le intitolazioni tramite un procedimento normato; dall'altro la spinta della società civile nell'indicare alcuni personaggi o eventi o concetti come importanti o identificativi per la comunità stessa. L'incontro tra questi due processi è raramente armonico, spesso oggetto di controversie e polemiche, in quanto la gestione della partecipazione dal basso è sovente trascurata rispetto alla salvaguardia dell'autonomia decisionale del governo comunale, non di rado intrisa di intenzionalità politica se non di vero e proprio intento propagandistico. In genere questo incontro tra le aspirazioni della comunità (che comunque non può intendersi come unitaria) e la volontà politica degli amministratori non solo non è efficacemente normato, ma soprattutto non è dotato di idonei strumenti che favoriscano il dialogo tra cittadino e istituzioni. Attualmente né la cittadinanza né gli enti amministrativi hanno contezza della stratificazione delle intitolazioni che il tempo ha depositato nel tessuto urbano, con la conseguente scarsa attenzione alle lacune maggiormente rilevabili in termini, ad esempio, di presenza femminile, o di rapporto tra personaggi civili, religiosi o militari, oppure del diverso rilievo degli eventi passati. Essendo invece l'odonomastica specchio di questa stratificazione e contemporaneamente messaggio - relativamente alle nuove intitolazioni - della visione continuamente cangiante e attuale della storia, avere uno strumento che consenta la consultazione, l'aggiornamento e la scelta degli odonimi costituirebbe un avanzamento di rilevo per la democrazia partecipativa [1, 3, 4] come anche per la ricerca storica e linguistica [11].

Al Laboratorio di Cultura Digitale dell'Università di Pisa stiamo affrontando il problema tenendo in considerazione, da un lato, gli strumenti del mondo delle Digital Humanities che potrebbero essere idonei alla sua risoluzione e, dall'altro, l'insieme delle questioni che pongono i dati dell'odonomastica. In questa proposta non si presentano strumenti già in avanzato stato di realizzazione, né soluzioni innovative: ci sembra invece interessante porre alla discussione della comunità degli umanisti digitali le sfide metodologiche che pone la stessa costruzione di un tale progetto di Digital Public History [10].

2. I DATI D'ORIGINE
Il primo passo per uno studio di questo patrimonio culturale materiale e immateriale è ovviamente quello di procurarsi

l'elenco degli odonimi del Comune di interesse. Ufficialmente esiste un Archivio Nazionale dei Numeri Civici e delle Strade Urbane (ANNCSU)[1], che dovrebbe essere gestito da ISTAT, in collaborazione con l'Agenzia delle Entrate. In pratica questa risorsa sembra essere consultabile solo dai Comuni, e, come privati cittadini siamo riusciti ad accedere soltanto alle specifiche tecniche di questo archivio. Una valida alternativa potrebbe essere OpenStreetMap (OSM)[2], che tramite una API permette di scaricare l'elenco degli odonimi relativi a una certa area geografica specificata nella richiesta, e quindi anche un Comune. Il problema di OSM è che gli elenchi scaricati contengono dati molto diversi per tipologia, in particolare presentano anche tutte le strade di accesso all'area richiesta, i nomi di "località" (frazioni, quartieri, aree speciali): diventa quindi molto difficile estrarre solo gli odonimi dell'area comunale. Un altro problema con OSM è che spesso gli odonimi presentano errori tipografici (ad es. Garobaldi invece di Garibaldi), il che rende ulteriormente complicata una mappatura automatica verso file di autorità.

Per il nostro lavoro siamo quindi partiti dall'elenco di odonimi ufficiale fornito dal Comune della Spezia, anche se presentava anch'esso - sebbene in misura minore - alcuni problemi: era riscontrabile una maggiore omogeneità nelle tipologie di dati rispetto a OSM, ma i termini apparivano in diverse varianti formali (nome di battesimo inesistente o espresso con lettera puntata o per esteso).

3. LA CATEGORIZZAZIONE E LA MATERIA OSCURA

Per una analisi socio-storica degli odonimi della Spezia, è stata definita una prima classificazione degli odonimi in quattro categorie (persone, luogo, evento/data, astratto/altro) e per ogni categoria sono state definite le caratteristiche principali, come mostrato nella tabella che segue:

<table>
<tr>
<td rowspan="2">persona
· genere: uomo | donna | collettivo
· religione: religioso | laico
· attività: artista, benefattore, imprenditore, militare, politico, sportivo, studioso
· rilevanza: locale | nazionale | internazionale
· vittima: SI | NO
· collocazione temporale: Età antica | Medioevo | Età Moderna | 1796-1915 | 1915-1943 | 1943-1946 | 1946-1992 | 1992-oggi
· riferimento tematico: Risorgimento ed età liberale | Fascismo e primo dopoguerra | Antifascismo e Resistenza | Lotta alla mafia | Anni di piombo | Lotte sociali e diritti umani | Altro</td>
<td>evento/data
· categoria: evento | struttura | ideale
· rilevanza: nazionale | internazionale
· collocazione temporale della persona o dell'evento ricordato: Età antica | Medioevo | Età Moderna | 1796-1915 | 1915-1943 | 1943-1946 | 1946-1992 | 1992-oggi
· riferimento tematico ossia il fenomeno storico che si intende celebrare o ricordare: Risorgimento ed età liberale | Fascismo e primo dopoguerra | Antifascismo e Resistenza | Lotta alla mafia | Anni di piombo | Lotte sociali e diritti umani | Altro</td>
</tr>
<tr>
<td>astratto/altro
· eventuale commento</td>
</tr>
</table>

La scelta delle categorie - in particolare di quelle relative alla collocazione temporale e al fenomeno storico richiamato (*riferimento tematico*) - è stata fatta in considerazione di quanto già emerso dagli studi sulle caratteristiche dell'odonomastica nazionale, che però non hanno mai proposto vere e proprie tipizzazioni generali, che prendessero in considerazione oltre che il genere e la professione anche il periodo storico e l'oggetto celebrato. Tutte le voci sono mutuamente esclusive, eccetto "attività" per la categoria "persona": unico campo dove è possibile inserire fino tre valori e dove si è scelto di limitare la granularità delle indicazioni. Così ad esempio poeti, scultori, pittori, scrittori e musicisti sono compresi nella categoria "artista". I fenomeni storici ricordati nelle intitolazioni sono, per ovvie ragioni, tarati sulla narrazione nazionale: una loro eventuale revisione sarà ovviamente possibile solo dopo aver concluso la classificazione di un *corpus* rappresentativo. Sono risultate scelte estremamente utili per lo studio dell'odonomastica italiana l'inserimento della categoria "vittima" per le persone e dell'attributo "locale" nella categoria geografica. Nel primo caso il "paradigma vittimario" è infatti ambito di studio recente e particolarmente interessante della storia contemporanea [8, 9] che sarebbe quindi utile riuscire a valutare numericamente in campioni significativi.
Data l'assegnazione prettamente manuale delle categorie "storiche" (*collocazione temporale e riferimento tematico*) abbiamo previsto un sistema di esclusione reciproca di tipologie impossibili perché contraddittorie, consentendo di volta in volta solo le tipologie possibili.

[1] ANNCSU Archivio nazionale dei numeri civici delle strade urbane, https://www.agenziaentrate.gov.it/portale/schede/fabbricatiterreni/portale-per-i-comuni/servizi-portale-dei-comuni/toponomastica.
[2] OSM OpenStreetMap, https://www.openstreetmap.org/.

Prima scelta Collocazione temporale	**Riferimento tematico permesso**		**Prima scelta Riferimento tematico**	**Collocazione temporale permessa**
Età antica			Risorgimento ed età liberale	1796-1915 tutti a seguire
Medioevo			Fascismo e primo dopoguerra	1915-1943 tutti a seguire
Età Moderna			Antifascismo e Resistenza	1943-1946 1946-1992 1992-oggi
1796-1915	Risorgimento ed età liberale		Lotta alla mafia	1946-1992 1992-oggi
1915-1943	Risorgimento ed età liberale Fascismo e primo dopoguerra		Anni di piombo	1946-1992 1992-oggi
1943-1946	Fascismo e primo dopoguerra Antifascismo e Resistenza		Lotte sociali e diritti umani	1946-1992 1992-oggi
1946-1992	Fascismo e primo dopoguerra Antifascismo e Resistenza Lotta alla mafia Anni di piombo Lotte sociali e diritti umani			
1992-oggi	Antifascismo e Resistenza Lotta alla mafia Anni di piombo Lotte sociali e diritti umani			

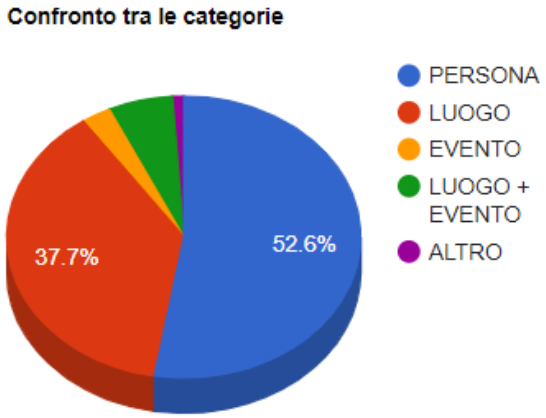


Figura 1. Rapporto tra le categorie principali negli odonimi della Spezia

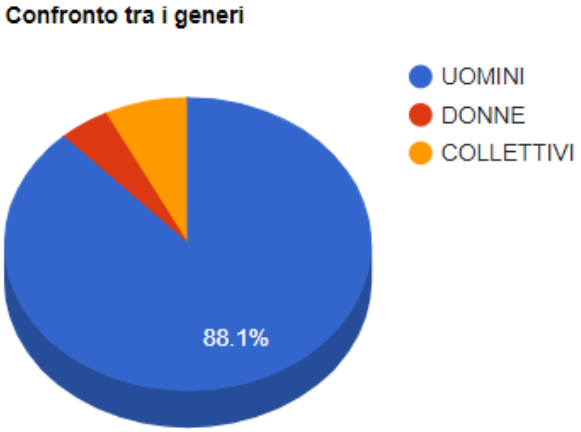


Figura 2. Rapporto tra uomini, donne e intitolazioni collettive negli odonimi della Spezia

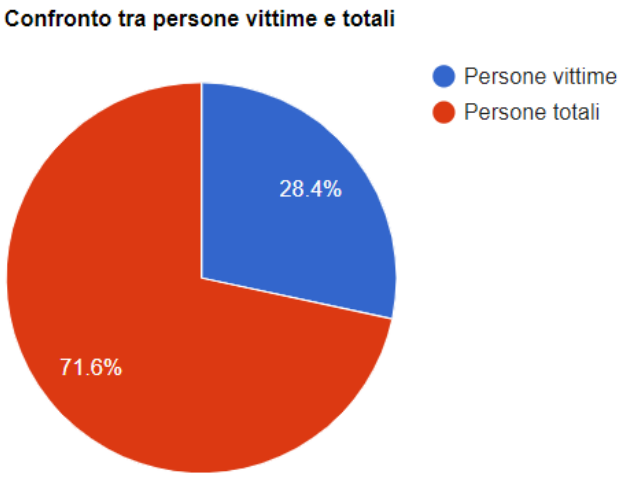


Figura 3. Percentuale delle vittime nelle intitolazioni a persone

Sono stati invece etichettati come "locali" tutti quegli odonimi che nascono originariamente come toponimi, ossia che non sono frutto di deliberazioni amministrative, ma costituiscono tracce residue di un periodo anteriore, di una nominazione spontanea avvenuta nel corso del tempo da parte di comunità con usi linguistici differenti (es. "della fornace", "della lobbia", Maggiano, Vecchiora, ecc.). Si tratta di termini che, per ovvie ragioni, non possono essere categorizzati come gli odonimi veri e propri e che in Italia sono presenti ovunque in percentuali ovviamente diverse, ma sempre rilevanti. Nel caso del database spezzino gli odonimi "locali" rappresentano il 65% dei luoghi e il 26% di tutti gli odonimi: una sorta di materia oscura sempre presente ma non analizzabile, perché non classificabile secondo le tipologie descritte e destinata a rispondere a differenti domande di ricerca.

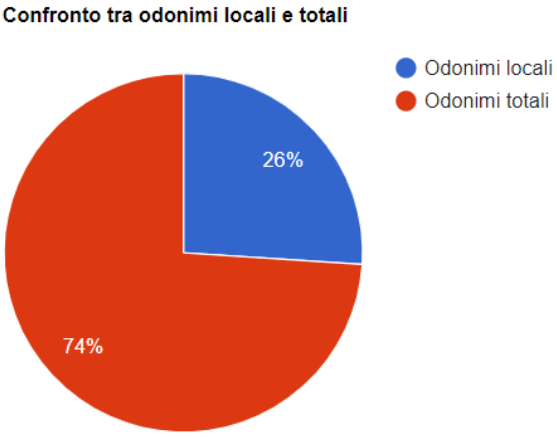


Figura 4. Rapporto tra toponimi e odonimi nel comune della Spezia

Come è facile intuire questa prima categorizzazione è stata fatta manualmente: anche se sarebbe stato possibile tentare una tipizzazione automatica per sesso o per appartenenza al dominio religioso (agiotoponimi), tutte le restanti annotazioni non potevano essere attribuite da un programma.
In aggiunta alle caratteristiche elencate sopra, di ogni odonimo viene anche verificata l'esistenza in Wikidata e in caso affermativo viene memorizzato il suo Wikidata ID (Qxxxxx) utilizzato a tutti gli effetti come file di autorità. Allo stato attuale, anche questo processo è essenzialmente manuale, dato il numero di definizioni che si possono trovare in Wikidata per la stessa entità ricercata. Anche aggiungendo l'informazione ulteriore di persona, luogo o evento, il numero di omonimi è infatti troppo elevato per essere gestito automaticamente. Inoltre, per persone (o eventi) la cui rilevanza sia strettamente locale, generalmente non si trova una corrispondente voce in Wikidata o, addirittura, si trovano omonimi, relativi ad altre entità in altre località. Relativamente a quest'ultimo problema, per inciso, riteniamo che il processo di schedatura e pubblicazione del database degli odonimi possa avere come benefica ricaduta l'arricchimento di Wikidata con le entità mancanti, fornendo una sorta di spinta virtuosa a riempire le pagine di Wikipedia a livello di comunità. Per gli odonimi della Spezia (circa 750), abbiamo identificato con certezza circa 400 entità di Wikipedia, tra locali e nazionali.

4. DA UNA CITTÀ ALLA NAZIONE

Nell'intento di creare un primo servizio utile alla comunità, ma soprattutto volendo in prospettiva allargare questa analisi socio-storica ad altri comuni italiani, stiamo sviluppando una applicazione *Web* che permetta di gestire agevolmente le informazioni relative ad ogni odonimo tramite un database relazionale (vd. Fig. 5) e che fornisca due distinti insiemi di funzionalità, usando gli odonimi del Comune della Spezia come banco di prova.

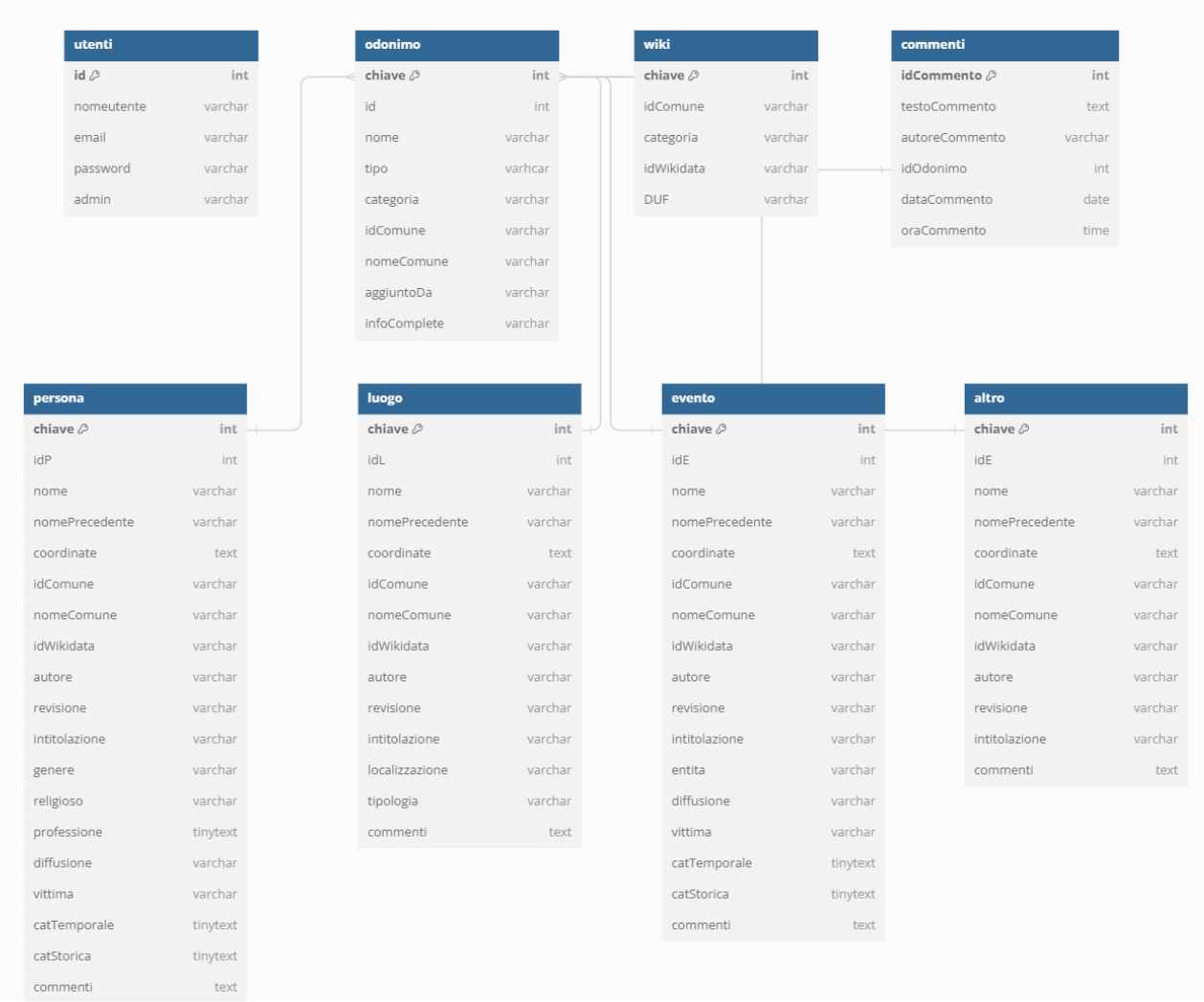


Figura 5. Schema logico del database

Un primo insieme di funzioni, a disposizione di tutti, permette di "navigare" attraverso gli odonimi già classificati, filtrandoli per nome o per caratteristica, con la possibilità di visualizzare in tempo reale i grafici delle diverse categorie e di scaricare in formato .csv i risultati della ricerca. Questo servizio - indubbiamente semplice dal punto di vista della ricerca nelle *Digital Humanities* - costituirebbe tuttavia uno strumento prezioso nell'ambito della *Public History*, in quanto fornirebbe ad amministratori e a cittadini uno spazio di confronto obiettivo e comune su cui valutare future decisioni sulla intitolazione degli spazi pubblici.

Il secondo insieme di funzioni, riservato agli utenti registrati, permette di definire o commentare le caratteristiche degli odonimi ancora non elaborati. Dopo aver effettuato il *login*, vengono mostrati gli odonimi già trattati da quell'utente e tutti quelli in attesa di definizione. L'utente può quindi scegliere di commentare o modificare gli odonimi da lui già definiti, oppure suggerire le caratteristiche di nuovi odonimi. Per gli utenti registrati è possibile aggiungere commenti ad ogni odonimo già classificato ed è possibile segnalare anche gli odonimi antichi, cancellati o mutati (fenomeno rilevante alla fine della Seconda Guerra Mondiale), dato che di solito questa informazione non è compresa nei database delle amministrazioni. È prevista infine una funzione di "super user", o amministratore, che può visualizzare i contributi e i commenti di tutti gli utenti per approvazione ed eventuali modifiche.

La classificazione degli odonimi della Spezia, e soprattutto il sistema sviluppato per facilitarne l'annotazione e il commento "dal basso", crediamo possano essere utilizzati per classificare anche gli odonimi di altre città, superando però almeno in parte il laborioso lavoro di categorizzazione manuale. Il salto di qualità potrebbe essere raggiunto, infatti, capitalizzando il lavoro già fatto al fine di automatizzare, in certa misura, la classificazione degli odonimi di una nuova città. Il "capitale" di partenza per questa automazione sarebbe l'insieme degli identificatori Wikidata, associati con sicurezza (tramite processo manuale) ai nomi di persona, o agli eventi o ai luoghi.
Il *dataset* Wikidata, oltre all'identificatore, contiene ovviamente anche il nome e il tipo (persona, luogo evento) dell'entità associata. Man mano che questo dataset viene costruito è possibile arricchirlo (parte manualmente e parte tramite uno *script*) con possibili varianti del nome dell'odonimo, ad esempio aggiungendo G. Garibaldi a Giuseppe Garibaldi. Questo faciliterà il riconoscimento del corretto ID Wikidata per odonimi provenienti da altri comuni. Prendendo inoltre in considerazione l'elenco degli odonimi di un'altra città con caratteristiche più o meno simili a quelle della Spezia (circa 100.000 abitanti) sarà infatti possibile identificare tutti gli odonimi già presenti nel *dataset* Wikidata e quindi assegnare loro i valori già utilizzati e validati da un "umano". Ripetendo questo processo per altre città, posizionate a diverse latitudini, ci si aspetta che il *dataset* di nomi associati a Wikidata continui ad aumentare, facilitando di conseguenza in maniera crescente l'aggiunta degli odonimi di nuovi comuni. Gli altri odonimi non presenti in Wikidata o facenti parte della "materia oscura", richiederanno invece sempre una classificazione manuale "dal basso", tramite l'utilizzo della medesima applicazione *Web* testata sul centro ligure. In quest'ultima attività, che possiamo definire di vero e proprio *crowdsourcing*, il *target* dei destinatari / utilizzatori non potrà che essere locale, mirato quindi sulla singola località, con conseguente costruzione di un'interfaccia *Web* studiata per facilitare al massimo la partecipazione del cittadino e personalizzata graficamente per agevolare l'identificazione.

In sostanza non riteniamo possibile, in virtù delle caratteristiche del patrimonio odonomastico italiano - costituito solo in parte da dati comuni e ripetitivi - avviare un progetto di annotazione condivisa a livello nazionale, in quanto i dati che sfuggono a ogni processo di automazione sono proprio quelli che appartengono alla comunità locale, che la comunità locale è in grado di riconoscere e anche valorizzare, ricordare e riconsiderare tramite il processo di riconoscimento proposto.

Solo quando sarà raggiunto per questo *dataset* un livello di confidenza accettabile, sarà infine possibile automatizzare quasi completamente il processo, fornendogli in ingresso il *dataset* di un comune e ottenendo in uscita il *dataset* degli odonimi riconosciuti e correttamente classificati, insieme al *dataset* (sperabilmente sempre più piccolo) degli odonimi da classificare manualmente. In questa fase probabilmente il problema maggiore sarà come ottenere per ogni comune un *dataset* di odonimi affidabile. In attesa che l'archivio ANNCSU diventi (facilmente) fruibile anche dai privati cittadini, la migliore alternativa è OSM, pur con tutte le limitazioni già dette.

5. CONCLUSIONI

Il progetto sulla schedatura partecipata degli odonimi d'Italia è ancora in una fase iniziale di studio, che qui viene portata proprio per uno, speriamo utile, confronto con la comunità scientifica. Riteniamo infatti importante aver spiegato gli intenti le problematiche che si incontrano nella categorizzazione degli odonimi del territorio nazionale e aver evidenziato difficoltà e opportunità nella scelta di lavorare su un livello locale rispetto a quello nazionale, anche e soprattutto in previsione del ricorso ineludibile alla partecipazione diretta del pubblico in attività di *Citizen Humanities* [5, 7].

BIBLIOGRAFIA

[1] Allegretti, Umberto. «Democrazia partecipativa e processi di democratizzazione». *Democrazia e diritto* fasc. II trimestre (2008): 1000–1043.

[2] Barzanti, Roberto. «Cultura e memoria nell'odonomastica stradale in Italia». *Bullettino senese di storia patria* CXXVI (2019): 550–55.

[3] Bobbio, Luigi. «Dilemmi della democrazia partecipativa». *Democrazia e diritto* IV trimestre (2006): 1000–1016.

[4] Castiglione, Mc, e M. Trovato. «Ricostruire una città, reinventare un'odonomastica». In *Visibile e invisibile: percepire la città tra descrizioni e omissioni*, 849–64. SCRIMM editore, 2014.

[5] Heinisch, Barbara. «Citizen Humanities as a Fusion of Digital and Public Humanities?» *Magazén* fasc. 2 (2020): JournalArticle_3442. https://doi.org/10.30687/mag/2724-3923/2020/02/001.

[6] Margotti, Marta. «Per le strade della patria. Nazionalizzazione e laicizzazione nell'odonomastica dell'Italia post-unitaria». *Rivista italiana di onomastica* 21, fasc. fasc. 2 (2015): 641–60.

[7] Paci, Deborah. «Conoscere è partecipare: digital public history, wiki e citizen humanities». *Umanistica Digitale* fasc. 10 (2021): 235–49. https://doi.org/10.6092/issn.2532-8816/12555.

[8] Ravveduto, Marcello. «I volti della Repubblica: il paradigma vittimario di Moro tra toponomastica, monumenti e anniversari civili». In *Aldo Moro, la storia e le memorie pubbliche*, 403:191–213. I libri di Viella, 2022.

[9] Ravveduto, Marcello. «Il paradigma vittimario della Repubblica: storia, memoria e media». *Ricerche Storiche* fasc. 3 (2022): 7–24.

[10] Ravveduto, Marcello, e Enrica Salvatori. «Storia digitale e digital public history: le novità di un antico mestiere». In *Digital Humanities: metodi, strumenti, saperi*, 229–54. Carocci, 2023.

[11] Samo, Giuseppe, e Francesco-Alessio Ursini. «Geographical maps meet place names where languages meet dialects: The case of Italian». *Forum Italicum: A Journal of Italian Studies* 57, fasc. 3 (novembre 2023): 1019–40. https://doi.org/10.1177/00145858231190030.

OpenStreetMap: uno strumento e uno spazio per la digital public history?

Camilla Zucchi
Università di Salerno, Italia - czucchi@unisa.it

ABSTRACT

La proposta ha l'obiettivo di coniugare estrazione e analisi dei dati provenienti dalla *neogeography* e dal *Volunteered Geographic Information* (VGI), con considerazioni storiografiche e approcci mutuati dalla *digital public history*, dal *cultural* e dal *monumental turn* e dalle *spatial humanities*. Si intende, dunque, riconsiderare OpenStreetMap alla luce delle sue potenzialità e del contributo che può apportare alla disciplina storica in termini di metodo, di fonti e di coinvolgimento del pubblico, facendo alcuni esempi pratici nel campo della monumentalistica e dell'odonomastica, temi oggi quantomai centrali nel dibattito pubblico.

PAROLE CHIAVE

Neogeography; Digital Public History; OpenStreetMap; Spatial Humanities; Monumental turn.

«Lo storico di domani o sarà un programmatore o non sarà» [24: 6]

1. INTRODUZIONE

Dal punto di vista della storia o meglio della geostoria, la piattaforma collaborativa di OpenStreetMap (OSM)[1] rappresenta una vera rivoluzione, ovviamente resa possibile dalla rivoluzione madre, il *digital turn*, e dal web 2.0 [7, 21, 39]. Se fino ai primi anni del secolo, per disegnare una mappa e disporre di informazioni geografiche erano necessari molti strumenti e precise competenze, oggi bastano un computer, una connessione internet e saper dove cercare i dati. Come Wikipedia si è imposta quale specchio e motore della *communis opinio* intorno a fatti e personaggi proprio perché si tratta di un'enciclopedia editabile da chiunque e ben ottimizzata, così OSM è diventato il punto di riferimento della *neogeography,* in cui, sin dalla sua creazione nel 2004 e in maniera sempre crescente, è risultato imprescindibile il *Volunteered Geographic Information (VGI)*, cioè una forma particolare di *User generated content* possibile grazie alla presenza di opportuni *plugin*. La *neogeography* ha, così, superato anche i GIS tradizionali, i quali richiedono ancora una capacità di programmazione significativa e approfondita[2]. La sua prima definizione, valida ancora oggi, descrive così la *neogeography*: «a diverse set of practices that operate outside, or alongside, or *in the manner of*, the practices of professional geographers. Rather than making claims on scientific standards, methodologies of neogeography tend toward the intuitive, expressive, personal, absurd, and/or artistic, but may just be idiosyncratic applications of "real" geographic techniques. This is not to say that these practices are of no use to the cartographic/geographic sciences, but that they just usually don't conform to the protocols of professional practice»[3]. Di fatto, la *neogeography* opera accanto ai geografi professionisti, senza alcuna pretesa su standard scientifici ma puntando all'utilizzo di mezzi veloci e facili, pur non aderendo ai protocolli disciplinari.
Su OSM, un'ampia sezione modello wiki serve per dare informazioni sui tag, sulle *keys* cui uniformare ciascun tag, sui diversi livelli o *layers* utilizzati all'interno della mappa globale[4] ed orientare così gli interessati a un uso standardizzato della piattaforma, al fine di ricavarne un risultato che risponda allo scopo originario: la libera circolazione, fruizione, inserimento ed eventuale correzione di dati geografici. La crescente fama dovuta all'enorme diffusione della piattaforma ha visto esplodere l'interesse intorno ad essa: secondo Muki Haklay, inoltre, i dati inseriti sono buoni e attendibili [20].

[1] https://www.openstreetmap.org/#map=10/45.6397/9.2740&layers=N cfr. [3, 9, 37: 106-117].

[2] Oggigiorno, però, la parola GIS «is taken as un umbrella term to describe the whole suite of software, standards, methods, tools, applications, standards and approches which allow the analysis and visualization of spatial relationships», anche se, specifica più avanti l'autore «GeoWeb […] refers to data or services which have geographical referents (such as a latitude and longitude) being published in a linkable way on a network, via a platform […]» [16: 14].

[3] Eisnor, Di-Ann. «Neogeography» *Platial,* https://web.archive.org/web/20060617110711/http://platial.typepad.com/news/neogeography/index.html. Wikipedia, voce *Neogeografia,* https://it.wikipedia.org/wiki/Neogeografia

[4] Wiki OpenStreetMap, voce *Main Page* https://wiki.openstreetmap.org/wiki/Main_Page

Una sempre maggiore disponibilità di dispositivi collegabili a internet permette un apporto più massiccio in termini numerici e più preciso in termini qualitativi delle informazioni geografiche, che spaziano dalla presenza di monumenti alle indicazioni stradali, da bar a ristoranti sino a negozi e stabilimenti balneari. Tutto ciò che è visibile trasmiga, così, dallo spazio fisico allo spazio virtuale, grazie al ruolo indispensabile degli *openstreetmappers*. Una comunità che è in grado di sfidare il servizio di Google Maps grazie alla natura *open source*: perché OSM, oltre a consentire l'inserimento dei dati, ne permette anche l'eventuale correzione o discussione in forum *ad hoc* e il download in un'API apposita. La cittadinanza, intesa in senso responsabile e attivo e qui identificata con gli *openstreetmappers*, viene continuamente coinvolta nel processo del fare storia: qui, infatti, ritorna la saldatura tra storia e geografia ed è con questi utenti che il *public historian* condivide l'autorità per capire e analizzare il presente spaziale visibile continuamente sulla mappa. Il *digital public historian* valuta dunque il «patrimonio culturale» aggiunto su OSM, non solo promuovendo la piattaforma ma provvedendo anche alla gestione della storia-memoria nello spazio pubblico, sia esso fisico o virtuale.

2. IL MONUMENTAL TURN

Quanto all'ambito di applicazione, negli ultimi decenni il *cultural turn* insieme al nuovo *spatial turn*[5] ha determinato un maggiore interesse storiografico e, soprattutto, nel dibattito pubblico intorno al campo delle rappresentazioni «visibili», tutto ciò che si connota come patrimonio pubblico condiviso[6], tanto che si è arrivati a parlare di un *monumental turn*. All'interno dell'ambito urbano, si tratta specialmente di monumenti e odonimi, che esprimono, o, talvolta, non esprimono più, l'identità collettiva di una città, di uno Stato e dei suoi abitanti: avere coscienza di come, quanto e da chi siano popolate le nostre città a livello di statue o di targhe viarie e di se e dove siano ancora visibili scomode eredità del passato può aiutare a generare, oltre a valide ricerche storiografiche, ipotesi di soluzione nella direzione dell'alterazione, della risemantizzazione, della musealizzazione, della rimozione o dell'integrazione, favorendo così un nuovo scambio dialogico di idee e opinioni. Il *cultural heritage* può divenire nel tempo fonte di dissidi se smette di essere in linea con il pubblico che ci interagisce [38]: elemento che è particolarmente forte oggi che «l'histoire s'écrit désormais sous la pression des mémoires collectives» [30: 302] e quindi ciascuna memoria collettiva porta avanti le proprie istanze e i propri simboli.

Seguendo questa direttrice, coniugare le neonate *spatial humanities* e la *digital public history* grazie alla *neogeography* apre indubbiamente importanti piste di studio, che aiutano a far luce a livello quantitativo e, secondariamente, a livello qualitativo su un ambito che, per quanto ora sia maggiormente studiato, risente della mancanza di un'indagine a tappeto e di uno standard metodologico di riferimento per essa. E la chiave di volta sta proprio nel ruolo da protagonista dell'utenza online, che contribuirebbe, consapevolmente proprio perché i dati OSM sono disponibili per il loro riutilizzo in base alla Open Database License (ODbL)[7], a una *citizen history* dello spazio condiviso [31].

La proposta, infatti, vorrebbe concentrarsi sullo studio di monumenti e odonimi italiani come parte del patrimonio storico e spia del fenomeno della nazionalizzazione politica, proprio per capire quali figure siano maggiormente rappresentate e individuarne, anche, l'area di più ampia diffusione e radicamento. Il risultato ambisce a essere un tentativo di destrutturare e leggere la stratificazione nelle nostre città di nomi e di personaggi che hanno costituito, e alcuni ancora costituiscono, l'identità nazionale e assegnarli ciascuno alla propria epoca di riferimento per arrivare a comprendere la complessità del presente. Si ritiene, dunque, di poter fornire un nuovo sostegno metodologico a un filone di studi che, senza un'analisi completa e in mancanza di database attendibili, si è dovuto limitare a contributi sì innovativi ma prettamente di impianto teorico, di carattere locale o focalizzati su singoli personaggi, determinati periodi o luoghi[8].

3. METODOLOGIA E OBIETTIVI

Ma, scendendo nel dettaglio a che cosa può essere utile OSM all'interno della *digital public history*?

In primis, appaiono del tutto evidenti il rapporto e la somiglianza tra il concetto di *VGI* e *crowdsourcing*: quest'ultimo viene definito come «[…] the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.» Seppur in assenza di un invito aperto, in questo caso così come per Wikipedia [37], l'utenza viene invogliata a contribuire a priori: sapendo le finalità del

[5] Da intendersi come l'esito degli sviluppi occorsi nelle «spatial technology – GIS, web mapping platforms, spatial data and infrastructure, digital, gazetteers and mobile dovices» [16: 41].

[6] A tal proposito rimangono imprescindibili gli studi già del 1978 di Maurice Agulhon, [1: 84-161].

[7] Wiki OpenStreetMap, voce *Open Database License* https://wiki.openstreetmap.org/wiki/Open_Database_License

[8] Per la bibliografia sull'odonomastica e la monumentalizzazione della storia, qualche riferimento d'insieme: [33], sulla scia dell'omologo francese: [10, 23, 26, 28]; sull'odonomastica e la monumentalistica dopo l'Unità: [11, 12: 376-380, 22: 330-349]; sull'odonomastica dopo la Prima Guerra Mondiale: [2, 6, 14, 18, 19]; sull'odonomastica e monumentalistica repubblicane: [4, 5, 8, 13, 15, 17, 25, 27, 29, 32, 34, 35, 36: 27-36].

progetto, gli iscritti, ad oggi più di 11 milioni nel mondo[9], provvedono ad inserire nuovi dati geografici e aggiornarli continuamente. Naturalmente, non tutti i continenti sono coperti con la stessa precisione, colpevole il *digital divide*. Ciononostante, l'approdo della rete in posti prima impensabili sta migliorando la mappatura. Tutti i dati, combinati e in vari formati processabili quali .geojson o .csv, possono essere scaricati attraverso un'API[10] con il linguaggio di programmazione come XML o uno *ad hoc* quale Overpass QL e, in seguito, analizzati.

La predisposizione di un codice in Overpass QL, più sintetico e preciso qui di XML, varia a seconda dell'oggetto: nel caso dell'odonomastica, la granularità della ricerca è stata maggiore, poiché la mole di dati è più ampia e quindi si scriverà una *query* per provincia; diversamente per i monumenti, dove si è lavorato seguendo aree più vaste, cioè le regioni. Premesso che gli elementi base di OSM sono *node* (punto), *way* (linea) e *relation* (relazione, cioè elemento che unisce gli altri elementi primari), cui possono essere aggiunti dei tag, e valgono sia per caricare che per scaricare i dati, per il *download* dei nomi strade e piazze l'operazione risulta più lineare tramite il ricorso a etichette in grado di rilevare ogni tipo di via, e cioè «residential or unclassified or tertiary or secondary or primary or service», specificando però che si vuole una sola occorrenza e non tanti nodi quanti costituiscono l'intera linea. Quanto ai tag che individuano i monumenti storici, possono essere molteplici e legati alla *key historic*: historic=monument; historic=memorial; historic=ruins; historic=military; oppure alla *key tourism* e presentare una specificazione legata al tag *artwork* e/o al tag *sculpture*, *bust* o *statue*. Ciascun utente, a regola, dovrebbe tracciare le opere pubbliche in base a precise caratteristiche spiegate nelle pagine wiki di riferimento, ma non è raro, ad esempio, che un *memorial* (memoriale), di norma più piccolo in dimensioni rispetto al monumento, venga taggato, invece, con *monument*. Altrettanto vale per *sculpture*, *bust* o *statue,* tra loro molto simili e per questo facilmente interscambiabili.

La ripulitura dei dati per calcolarne la frequenza è stata effettuata attraverso Excel, Openrefine e alcuni comandi da terminale.

Il seguente intervento vorrebbe, dunque, utilizzare, a titolo esemplare, la *digital public history*, sostenuta e integrata da riflessioni intorno alla *neogeography* e dall'approccio *VGI*, per (ri)unire storia e geografia con un'attenzione primaria alle rappresentazioni pubbliche collettive, alcune oggi così controverse e dibattute, per fornire uno standard metodologico e da lì partire per considerazioni da *public historian*.

BIBLIOGRAFIA

[1] Agulhon, Maurice. «Histoire vagabonde». In *Ethnologie et politique dans la France contemporaine*, 1:84–161. Paris: Éditions Gallimard, 1988.

[2] Albanese, Giulia. «Mappare la memoria del fascismo». In *I luoghi del fascismo. Memoria, politica, rimozione*, a cura di Lucia Ceci e Giulia Albanese, 31–54. Roma: Viella, 2022.

[3] Arsanjani, Jamal Jokar, Alexander Zipf, Peter Mooney, e Marco Helbich, (a cura di). *OpenStreetMap in GIScience. Experiences, Research, and Applications*. Heidelberg-New York-London: Springer, 2015.

[4] Baioni, Massimo. «Demolire il littorio. Tragitti della simbologia fascista nell'Italia repubblicana». *Memoria e Ricerca* 63, fasc. 1 (2020): 181–94.

[5] Baioni, Massimo. «Le Fascisme italien entre histoire et mémoire. Le problème du musée à Predappio». *Synergies Roumanie* 15 (2020): 15–22.

[6] Balzani, Roberto. «Urban Toponymy, Cultural Memory and the World Wars». In *Memories and Representations of War. The Case of World War I and World War II*, a cura di Elena Lamberti e Vita Fortunati, 89–102. Amsterdam-New York: Rodopi, 2009.

[7] Batty, Michael, Andrew Hudson-Smith, Richard Milton, e Andrew Crooks. «Map Mashups, Web 2.0 and the GIS Tevolution». *Annals of GIS* 16, fasc. 1 (2010): 1–13. https://doi.org/10.1080/19475681003700831.

[8] Belmonte, Carmen. *A Difficult Heritage. The Afterlives of Fascist-era Art and Architecture*. Cinisello Balsamo: Silvana Editoriale, 2024.

[9] Bennett, Jonathan. *OpenStreetMap. Be Your Own Cartographer*. Birmingham: Packt Publishing, 2010.

[10] Berggren, Lars, e Lennart Sjöstedt. *L'ombra dei grandi. Monumenti e politica monumentale a Roma (1870-1895)*. Roma: Artemide, 1996.

[11] Bracco, Barbara. «Tendenze educative e istanze politiche della classe dirigente milanese: i luoghi dell'identità nazionale nella toponomastica del capoluogo lombardo dall'Unità alla Grande guerra». In *Riforme e istituzioni fra Ottocento e Novecento*, a cura di Luigi Cavazzoli e Carlo G. Lacaita, 395–426. Manduria: Lacaita, 2001.

[12] Brice, Catherine. «*Monarchie et identité nationale en Italie: 1861-1900*». Paris: Éditions de l'École des hautes études en sciences sociales, 2010.

[9] Wiki OpenStreetMap, voce *Users and GPX Uploads* https://wiki.openstreetmap.org/wiki/Stats#Users_and_GPX_uploads
[10] https://overpass-turbo.eu/

[13] Brogi, Daniela. *Lo spazio delle donne*. Torino: Einaudi, 2024.

[14] Castelnovi, Michele, e Arturo Gallia. «Geografia della memoria odonomastica della Grande Guerra». *Bollettino della Società geografica italiana* 9, fasc. XIII (2009): 431–46.

[15] Cazzato, Vincenzo. *Natura aere perennius. Parchi della Rimembranza e luoghi della memoria*. Ravenna: Montanari Editore, 2022.

[16] Dunn, Stuart. *A History of Place in the Digital Age*. London-New York: Routledge, 2019.

[17] Focardi, Filippo. «Ricordare il passato. Usi pubblici della storia e della memoria in Italia dopo la prima Repubblica». In *Riparare, Risarcire, Ricordare. Un dialogo tra storici e giuristi*, a cura di Giorgio Resta e Vincenzo Zeno-Zencovich, 241–72. Napoli: Editoriale Scientifica, 2012.

[18] Gallia, Arturo. «Cartografia storica e strumenti digitali per lo studio della memoria della grande guerra. L'odonomastica capitolina». In *La Grande Guerra. Luoghi, eventi, testimonianza, voci*, a cura di Simonetta Conti, 311–28. Canterrano: Aracne, 2018.

[19] Giadrossi, Andrea. «Leggi razziali e odonomastica a Trieste». *Qualestoria* 2 (2012): 117–34.

[20] Haklay, Muki. «How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets». *Environment and Planning B: Planning and Design* 37 (2010): 682–703.

[21] Hudson-Smith, Andrew, Andrew Crooks, Maurizio Gibin, Richard Milton, e Michael Batty. «NeoGeography and Web 2.0: concepts, tools and applications». *Journal of Location Based Services* 3, fasc. 2: NeoGeography (2009): 118–45.

[22] Isnenghi, Mario. *Le guerre degli italiani. Parole, immagini, ricordi 1848-1945*. Bologna: Il Mulino, 2005.

[23] «Le città leggibili. La toponomastica urbana tra passato e presente. Atti del convegno di studi (Foligno, 11-13 dicembre 2003)», Vol. 2004. Perugia: Deputazione di storia patria per l'Umbria, 2004.

[24] Le Roy Ladurie, Emmanuel. *Le frontiere dello storico*. Roma-Bari: Laterza, 1976.

[25] Mask, Deirde. *Le vie che orientano. Storia identità e potere dietro i nomi delle strade*. Torino: Bollati Boringhieri, 2021.

[26] Mastrelli, Carlo Alberto, (a cura di). *Odonomastica. Criteri e normative sulle denominazioni stradali. Atti del convegno Trento 25 settembre 2002*. Trento: Nuove Arti Grafiche, 2005.

[27] Messina, Dino. *La storia cancellata degli Italiani*. Milano: Solferino, 2022.

[28] Milo, Daniel. «Les noms des rues». In *Les lieux de mémoire*, II: La Nation, sous la direction de Pierre Nora:1887–1918. Paris: Éditions Gallimard, 1997.

[29] Montanari, Tomaso. *Le statue giuste*. Roma-Bari: Laterza, 2024.

[30] Nora, Pierre. *Présent, nation, mémoire*. Paris: Éditions Gallimard, 2011.

[31] Paci, Deborah. «Conoscere è partecipare: digital public history, wiki e citizen humanities». *Umanistica Digitale* 10 (2021). https://doi.org/10.6092/issn.2532-8816/12555.

[32] Parola, Lisa. *Giù i monumenti? Una questione aperta*. Torino: Einaudi, 2022.

[33] Raffaelli, Sergio. «I nomi delle vie». In *I luoghi della memoria. Simboli e miti dell'Italia unita*, a cura di Mario Isnenghi, 215–42. Roma-Bari: Laterza, 1996.

[34] Ravveduto, Marcello. «La toponomastica della Seconda Repubblica. Falcone e Borsellino, vittime della mafia». *Memoria e Ricerca* 57, fasc. 1 (2018): 157–74.

[35] Ridolfi, Maurizio. «Il nuovo volto delle città. La toponomastica negli anni della transizione democratica e della nascita della Repubblica». *Memoria e Ricerca* 20, fasc. 3 (2005): 147–67.

[36] Ridolfi, Maurizio. *Verso la public history. Fare e raccontare storia nel tempo presente*. Pisa: Pacini Editore, 2017.

[37] Romano, Antonello. *La geografia delle piattaforme digitali. Mappe, spazi e dati dell'intermediazione digitale*. Firenze: Firenze University Press, 2022.

[38] Testi, Arnaldo. *I fastidi della storia. Quale America raccontano i monumenti*. Bologna: Il Mulino, 2023.

[39] Warf, Barney, e Daniel Sui. «From GIS to Neogeography: Ontological Implications and Theories of Truth». *Annals of GIS* 16, fasc. 4 (2010): 197–209. https://doi.org/10.1080/19475683.2010.539985.

Preservare la memoria: il progetto Storage e l'archivio dell'ex istituto di Archeologia

Giovanni Fragalà,[1] Pietro Maria Militello[2]
[1]CNR Istituto di Scienze del Patrimonio Culturale, Catania, Italia - giovanni.fragala@cnr.it
[2]Università di Catania, Italia - milipi@unict.it

ABSTRACT[1]
L'intervento presenta i risultati di un progetto di ricerca finanziato dall'Ateneo di Catania. Tra le attività avviate, si presenta il caso specifico dell'archivio fotografico dell'ex Istituto di Archeologia dell'Università di Catania. La documentazione fotografica ha costituito il punto di avvio per la elaborazione di una schedatura e la creazione di una piattaforma per rendere fruibile il materiale sul web, nell'ottica della Scienze Aperta. Nello stesso momento ha costituito un momento di riflessione sulla gestione di classi di dati apparentemente molto diverse tra di loro come manufatti archeologici, documenti di archivio, materiale fotografico.

PAROLE CHIAVE
Archives; Archaeology; Photography; Database.

1. INTRODUZIONE
L'Archivio dell'ex Istituto di Archeologia, conservato presso la sede di Palazzo Ingrassia del Dipartimento di Scienze Umanistiche, conserva diversi materiali provenienti dai vari gabinetti ed istituti di Archeologia che si sono susseguiti presso l'Ateneo catanese fino al 1999, anno della istituzione dei dipartimenti [5, 12, 13]. Esso conserva strumentazioni per la riproduzione audiovisiva, materiale didattico, tesi di laurea, reperti archeologici a scopo didattico e soprattutto una grande quantità di materiale fotografico comprensivo di diapositive, negativi, microfiches (v. infra).
Nell'ambito di una attività di catalogazione del materiale, resa necessaria dal suo trasferimento dalla collocazione originaria in Via di Sangiuliano 269 all'attuale in Via Biblioteca 4, si è posto già per tempo il problema delle modalità di catalogazione da utilizzare tenendo conto di due, opposte, esigenze: quella di garantire una adeguata catalogazione a materiali di tipo molto diverso e quella di mantenere una unità nella gestione dell'archivio come entità culturale, formatasi cioè a seguito di esigenze convergenti di tipo didattico e scientifico. Per il primo aspetto, se indicazioni come collocazione e dimensione sono comuni, altre voci sono specifiche di ciascuna classe ed apparentemente incompatibili: indicazioni sullo spessore o il peso sono necessarie nel caso del materiale archeologico, ma non negli altri materiali; la voce "contesto di provenienza" di un manufatto archeologico non è presente nella classificazione di una tesi di laurea o di un documento amministrativo, e per converso di un manufatto non si chiede l'autore (o la ditta), come nel caso di una fotografia. La "cronologia" ha significato diverso per un manufatto e per uno strumento. E tuttavia, e qui veniamo al secondo aspetto, questi materiali così eterogenei hanno fatto parte di una realtà unitaria, quella dell'Istituto di Archeologia, di cui restituiscono la storia [1]: informazioni sulla data di ingresso di un documento fotografico acquistato potrebbero essere illuminanti per chiarire scelte didattiche, come l'attivazione di un nuovo corso o l'introduzione di nuove modalità di insegnamento, così come l'ingresso di materiale archeologico può essere messo in rapporto con le attività di ricerca sul campo.
Tenendo presenti queste esigenze, gli obiettivi dell'attività sono stati due: rendere disponibili i dati on-line e utilizzare una piattaforma in cui classi di dati differenti potessero essere analizzati ed interrogati in maniera unitaria, contemporaneamente.

2. LA SEZIONE FOTOGRAFICA DELL'EX ISTITUTO DI ARCHEOLOGIA
Il nucleo fotografico rappresenta il principale componente della documentazione dell'Archivio. Esso possiede non solo un valore intrinseco, ma anche storico: le fotografie documentano in gran parte gli scavi e le ricerche archeologiche effettuate nel territorio dall'Istituto di Archeologia dell'Ateneo catanese, i reperti conservati presso musei, nonché fotografie tratte da pubblicazioni archeologiche e di storia dell'arte fatte realizzare con finalità didattiche a partire dagli inizi del Novecento. Si tratta di materiale di grande interesse per lo studio di differenti aspetti legati non solo alla ricostruzione delle strumentazioni e delle tecniche fotografiche con cui le immagini furono realizzate, ma anche alla diffusione della fotografia come strumento di supporto alla didattica e alla ricerca universitaria, nonché alla determinazione di quelle personalità, sia docenti che tecnici specializzati, che vollero la costituzione dell'Archivio e ne resero possibile il suo costante incremento

[1] L'introduzione e la conclusione sono di P. M. Militello, mentre a G. Fragalà sono dovuti i paragrafi centrali.

nel corso degli anni.
Ultimo aspetto, non meno importante degli altri, riguarda il valore documentale insito nel materiale fotografico in relazione ai contesti rappresentati, di cui spesso testimoniano lo stato di conservazione al momento della scoperta o, in alcuni casi, prima che interventi successivi ne alterassero in vario modo la natura o ne determinassero l'irreparabile perdita.
Lo studio sinora condotto ha consentito di isolare all'interno dell'archivio quattro nuclei, differenti per tipologia di materiali e per quantità. Il fondo storicamente più significativo, che è stato l'oggetto principale di questa ricerca, comprende i negativi in bianco e nero su lastra di vetro, ottenuti applicando il procedimento alla gelatina bromuro ai sali d'argento. Esso risulta composto da 3242 fototipi in diversi formati, che vanno dal 6x9 al 18x24. A cui devono essere aggiunte 1133 diapositive in bianco e nero su lastra di vetro nel formato 8x8 e 9,5x10.
Un secondo gruppo è composto di circa 50.000 negativi in bianco e nero su pellicola, in cui sono presenti i formati 6x9, 6x6 e 35 mm. Il terzo fondo raccoglie oltre diecimila diapositive in bianco e nero e a colori su pellicola in poliestere e un quarto fondo è costituito da positivi fotografici ottenuti in varie tecniche, dalle albumine alle stampe alla gelatina ai sali d'argento, e diversi album fotografici [5, 6].

Figura 1. Un gruppo di lastre fotografiche in fase di riordino

La digitalizzazione e soprattutto la classificazione di questi documenti non è stata esente da problemi. La schedatura dei materiali, quando esistente, non ha seguito, come in molte altre istituzioni, un progetto organico e ben strutturato, che ne ha fin dall'inizio regolato l'evoluzione guidando sia le fasi dell'acquisto di collezioni di immagini o di attrezzature, sia quelle della realizzazione al suo interno di nuova documentazione. L'archivio, piuttosto, si è formato progressivamente, per stratificazione naturale, in risposta a situazioni contingenti, legate ora ad esigenze di didattica, ora a necessità di documentazione delle attività di ricerca svolte. Inoltre non sempre sono annotate su ogni fototipo quelle informazioni che oggi sarebbe estremamente utile possedere, come la data, il luogo della ripresa, il soggetto, oppure l'autore; ad aggravare la perdita di informazioni preziose subentra, inoltre, l'abitudine, molto diffusa, a inventariare e ad archiviare, in modo sistematico, i materiali fotografici solo a distanza di anni dalla loro realizzazione.
Per questo motivo, è stato necessario far precedere lo studio da un riordino del patrimonio fotografico esistente (vd. Fig. 1) e contestualmente a esso da una catalogazione integrale, l'una e l'altra accompagnate da una raccolta, analisi e interpretazione delle fonti documentali disponibili, necessaria non solo per ricostruire la storia della sua formazione, ma anche per fondare su solidi e oggettivi criteri storici il processo di sistemazione e catalogazione avviato. Si sono riordinate le lastre e le fotografie sulla base dei differenti formati e, per ciascuno di essi, in base all'ordine progressivo che era stato a suo tempo assegnato. L'analisi ha consentito, anche in assenza dell'inventario originario, che non è stato possibile reperire, di ripristinare il criterio adottato nella prima fase di catalogazione e archiviazione dei materiali. Questo prevedeva l'attribuzione a ciascuna lastra fotografica di un codice formato da un numero romano che identificava il formato della

lastra, seguito da un numero arabo che indicava l'ordine cronologico di acquisizione[2].
Durante la sistemazione, è risultato evidente come a volte la progressione numerica per lo stesso soggetto presentasse dei vuoti anche di diverse decine di unità, che potrebbero tradire un processo di archiviazione e catalogazione non sempre contestuale all'acquisizione del singolo esemplare e condotto a termine subito dopo lo sviluppo. Anzi, è probabile che tutte le lastre fotografiche siano state catalogate a distanza di anni proprio con l'intento di riorganizzare l'archivio fotografico. In mancanza di indicazioni precise nella documentazione rinvenuta, è possibile, in via ipotetica, supporre che la catalogazione e prima sistemazione dei materiali sia avvenuta nel momento in cui i due istituti di Archeologia e di Storia dell'Arte si siano separati.

3. DATA STORAGE: LA DIGITALIZZAZIONE DELLE INFORMAZIONI

Contestualmente a questa prima ricognizione e alla pulizia dei materiali è stata avviata una campagna di digitalizzazione finalizzata all'acquisizione delle lastre negative, delle diapositive su vetro e di una parte del materiale positivo[3].
La digitalizzazione delle lastre è stata eseguita ponendo i negativi sopra un piano traslucido retroilluminato con luce led a bassa emissione di raggi UV. Ogni singola lastra è stata in tal modo fotografata con una fotocamera ad alta risoluzione (*full frame*) montata su uno stativo. Oltre alle lastre negative sono state digitalizzate un cospicuo numero di diapositive (vd. Fig. 2), risalenti in gran parte agli anni Venti del Novecento, realizzate e commercializzate da Franz Stoedtner, uno dei pionieri della editoria fotografica (1870 -1946).

Figura 2. Diapositive edite dalla Franz Stoedtner di Berlino. Archivio fotografico storico ex Istituto di Archeologia.

L'intero processo è stato preceduto dalla definizione di alcune linee guida finalizzate all'adozione di una metodologia di acquisizione. Questa si è fondata sul riesame sia delle normative ministeriali emanate dall'Istituto Centrale per il Catalogo e la Documentazione (ICCD)[4] e dall'Istituto Centrale per il Catalogo Unico (ICCU), sia su un'analisi delle esperienze maturate, da chi scrive, in seno ai più importanti archivi nazionali[5] [2, 3; 4, 7, 11, 13, 14].
Per il trattamento archivistico si è scelto di utilizzare la piattaforma di catalogazione xDams[6] rilasciata con licenza *open source*. Si tratta di un software elaborato da regesta.exe grazie ad un progetto europeo dal titolo "Digital Archives and Memory Storages" (2000-2004) e continuamente aggiornato. Tale software è scaricabile in versione GNU GPL3, una licenza che permette l'utilizzo, la modifica e la condivisione del codice sorgente, a patto di rilasciare eventuali interventi e modifiche alle medesime condizioni.

[2] Con il numero I veniva indicato il formato di lastra 18x24; III per il 13x18; V per il 10x15; con VII il formato 9x12 e, infine, con IX le lastre 6x9.
[3] http://www.iccd.beniculturali.it/it/1/home
[4] Mibac (Ministero per i Beni e le attività culturali). Istituto Centrale per il Catalogo e la Documentazione. *Normativa per l'acquisizione digitale dele immagini fotografiche.* Roma 1998.
[5] Mandolesi, Sveva. "xDams. Tracciati archivio fotografico". Luglio 18, 2014. http://xdams.org/supporto/documentazione
[6] www.xdams.org

L'applicativo ha mutuato e adattato, semplificando, il modello dati per la descrizione dei beni fotografici della scheda "F" (Fotografia) elaborata nel 1999 dal Ministero per i Beni e le Attività Culturali, attraverso l'Istituto Centrale per il Catalogo e la Documentazione e giunta ora alla versione aggiornata 4.0 al gennaio 2024[7].

Il catalogo fotografico è stato organizzato con una struttura gerarchica, che fa capo a una scheda "radice" principale, denominata "Archivio fotografico storico Disum". Per la strutturazione dell'archivio si è adottato un criterio di ordine contenutistico, che fa riferimento ad una macro classificazione degli oggetti o dei contesti rappresentati. Tale principio ha consentito di isolare 28 schede "madri", corrispondenti ad altrettante classi archeologiche [5].

Per ciascuna scheda madre sono state create diverse schede "figlie" di livello inferiore deputate ciascuna all'archiviazione di *dataset* relativi ad aspetti peculiari (vd. Fig. 3). Fra queste un posto di rilievo occupa quella deputata alla descrizione analitica delle singole fotografie. La struttura di quest'ultima scheda costituisce, come detto, una versione semplificata del tracciato della scheda "F" elaborata dall'ICCD ed è articolata su cinque sezioni: "identificazione"; "descrizione"; "status della documentazione"; "status delle note" e "compilazione" (2; 12). In queste sezioni vanno inserite tutte le informazioni relative agli aspetti fisici e contenutistici di ciascun documento archiviato. Una volta compilati i campi, la piattaforma consente di navigare direttamente attraverso l'albero gerarchico oppure di effettuare ricerche libere o per voci specifiche (vd. Fig.4).

Figura 3. Esempio di struttura gerarchica realizzata con xDams

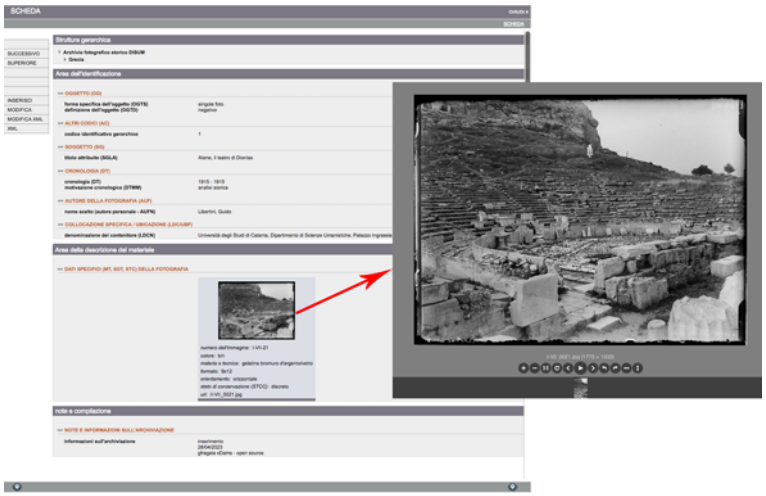


Figura 4. Esempio di scheda sintetica di ricerca

[7] http://www.iccd.beniculturali.it/it/ricercanormative/62/f-fotografia-4_00

4. DAI DATI AL WEB: LA CREAZIONE DI UNA PIATTAFORMA DI CONDIVISIONE

Uno degli obiettivi del progetto di ricerca è stato quello di creare le condizioni per consentire al maggior numero di utenti di accedere al patrimonio fotografico e alle informazioni che lo riguardano e che contribuiscono a una piena comprensione dei suoi valori storici e documentali. L'applicativo xDams permette di conservare, organizzare, condividere e valorizzare singoli patrimoni archivistici e, essendo stato sviluppato in modalità ASP (*Application Service Provider*), è interamente *Web based*. Ciò consente di utilizzare il canale web per l'accesso alle informazioni e il linguaggio XML (*eXtensible Markup Language*) per la conservazione e condivisione dei dati.

La consultazione può avvenire tramite l'apposita interfaccia, che consente una ricerca libera o strutturata attraverso l'indicazione di specifici criteri di interrogazione. A questa modalità si aggiunge una navigazione basata su un modello gerarchico ad "albero". Quest'ultima consente la visualizzazione rapida delle risorse sotto forma di scheda "sintetica" contenente le informazioni più importanti, mentre l'esplorazione completa di tutte le informazioni è demandata ad un'apposita scheda "estesa". La piattaforma permette, inoltre, la gestione dell'archivio e l'organizzazione delle risorse mettendo a disposizione funzioni di ordinamento automatico e di editing dei record ("taglia", "copia" e "incolla"). Oltre la consultazione in modalità intranet, i dati possono essere pubblicati su siti e proposti con canali tematici di fruizione grazie alla possibilità di associare interfacce di comunicazione tra la base dati e il *front-end web*.

La catalogazione è ancora in corso, pertanto, nello svolgimento del progetto ci si è limitati alla strutturazione, al testing e alla valutazione della piattaforma, commisurata agli obiettivi previsti e ai risultati attesi. Ciò è avvenuto attraverso l'inserimento di dati campione finalizzati alla verifica complessiva delle funzionalità del sistema, auspicando che future attività possano portare al completamento della catalogazione avviata e la sua accessibilità online tramite un apposito sito web. La piattaforma è risultata perfettamente confacente alle esigenze del progetto.

Nello specifico per la sperimentazione della piattaforma xDams nell'ambito del progetto è stato attivato un account gratuito con uno spazio di archiviazione di quattro gigabyte, che ha permesso di testarne e valutarne le funzionalità e di gestire, anche se non in forma completa, una prima fase di implementazione dell'archivio.

5. CONCLUSIONI

Il lavoro di catalogazione e digitalizzazione del patrimonio fotografico dell'ex Istituto di Archeologia ha permesso di testare la piattaforma xDams che si configura come uno strumento ideale per la digitalizzazione di archivi complessi, come quello qui presentato. Sebbene in questo caso i materiali da catalogare appartengono a differenti tipologie (strumenti per la riproduzione audiovisiva, materiale didattico, tesi di laurea, reperti archeologici a scopo didattico, materiale fotografico, ecc.) bisogna tenere in considerazione che da un punto di vista informatico alcune di queste differenze tendono ad annullarsi. Infatti, una digitalizzazione ottimale dei dati necessita di una loro preliminare normalizzazione seguendo il principio della minimalità, al fine di limitare al minimo le ridondanze, il che comporta una standardizzazione non solo terminologica, ma soprattutto concettuale [8, 9]. Nel caso specifico, informazioni quali "spessore" e "peso", valorizzabili solo per alcune classi di materiali, concettualmente si possono inserire in una entità generica "dimensioni" atta al trattamento di qualsivoglia informazione dimensionale; un archivio che gestisce i dati relativi alla "datazione" può trattare informazioni cronologiche inerenti a un reperto archeologico, ma anche datazioni relative a documenti o strumentazione; e così via. Fondamentale, in tal senso, è la flessibilità del software che, permettendo la modifica del codice sorgente, può essere piegato alle esigenze specifiche dei casi studio. Il vantaggio, quindi, è duplice: da un punto di vista informatico la gestione in un unico applicativo di differenti dataset permette interrogazioni trasversali, in grado di mettere in evidenza associazioni contestuali e collegamenti fra le informazioni altrimenti impercettibili, oltre a permettere una lettura inedita dei dati, dal punto di vista archivistico si mantiene l'entità culturale unitaria dell'archivio.

6. RINGRAZIAMENTI

La ricerca è stata condotta nell'ambito del progetto interdipartimentale (DISUM-DMI) "Storage. Dai dati al Web", finanziato nell'ambito del programma Pia.Ce.Ri dell'Università di Catania (2021-2024). Lo studio, la classificazione e digitalizzazione dell'archivio fotografico si è svolta nell'ambito della tesi di ricerca dottorale di Giovanni Fragalà [10].

BIBLIOGRAFIA

[1] Abbattista, Guido. «Risorse elettroniche e telematiche per gli studi di Storia moderna». *Memoria e Ricerca* 8, fasc. 6 (2000): 177–87.

[2] Berardi, Elena. *Normativa F - Fotografia versione 4.00 Strutturazione dei dati e norme di compilazione*. Roma: Istituto Centrale per il Catalogo e la Documentazione, 2016.

[3] Berselli, Silvia, e Laura Gasparini. *L'archivio fotografico. Manuale per la conservazione e la gestione della fotografia antica e moderna*. Bologna: Zanichelli, 2000.

[4] Bieman, Ern, (a cura di). *Capture Your Collections: A Guide for Managers Who Are Planning and Implementing Digitization Projects*. Ottawa: Canadian Heritage Information Network, 2020.

[5] Bramante, Daniela Maria. «La classificazione e il riordino dell'archivio fotografico dell'ex Istituto di archeologia». In *Interferenze. Un dialogo tra scienze umane e scienze dure*, a cura di Figuera Marianna, 137–45. Catania, 2016.

[6] Buscemi, Francesca. «Fotografia e Archeologia. L'archivio presso il dipartimento di Studi Umanistici dell'Università di Catania». *Agorà* aprile – giugno (2011): 36–40.

[7] Falchetta, Piero. «Guida breve alla digitalizzazione in biblioteca». *Biblioteche oggi* 9 (2000): 52–67.

[8] Figuera, Marianna. (a cura di). *Interferenze. Un dialogo tra scienze umane e scienze dure*. Catania: Corso internazionalizzato in archeologia, 2016.

[9] Figuera, Marianna. *Un sistema per la gestione dell'affidabilità e dell'interpretazione dei dati archeologici. Percezione e potenzialità degli small finds: il caso studio di Festòs e Haghia Triada*. Vol. 8. Praehistorica Mediterranea. Oxford: Archaeopress, 2020.

[10] Fragalà, Giovanni. «Fotografia e archeologia nell'Ateneo catanese. Tesi di dottorato (XXXV ciclo)». Università di Catania, 2020.

[11] Frey, Franziska S., e James M. Reilly. *Digital imaging for photographic collections. Foundations for technical standards*. Rochester: Image Permanence Institute Rochester Institute of Technology, 1999.

[12] Militello, Pietro M. «Immagini e strumenti. L'archeologia catanese attraverso l'archivio fotografico». In *Interferenze. Un dialogo tra scienze umane e scienze dure*, a cura di Marianna Figuera, 133–36. Catania, 2016.

[13] Militello, Pietro M., Salvo Adorno, e Anna Maria Seminara. «L'archeologia nell'università di Catania: pratica della didattica e Terza Missione nel secondo dopoguerra». In *L'Archeologia in Sicilia nel secondo dopoguerra. Atti convegno Catania*, a cura di Fabrizio Nicoletti e Rosalba Panvini, 87–101. Palermo, 2020.

[14] Ostrow, Stephen E. *Digitizing Historical Pictorial Collections for the Internet*. Alexandria, USA: Council on Library and Information Resources, 1998.

Un Atlante digitale per la storia marittima del Regno di Sardegna

Giampaolo Salice
Università degli Studi di Cagliari, Italia – giampaolo.salice@unica.it

ABSTRACT

L'intervento è diretto a presentare l'Atlante digitale per la Storia marittima della Sardegna (ASMSA), strumento impiegato per la costituzione di un quadro conoscitivo complessivo della storia marittima sarda con attenzione specifica all'età moderna. L'Atlante integra la bibliografia di riferimento, sia con la documentazione custodita in archivi locali, nazionali e internazionali, sia con dati e applicativi generati dal campo, anche con azioni di public engagement. L'intervento illustrerà la piattaforma tecnologica e il flusso di lavoro attraverso le quali i dati, raccolti secondo regole condivise, vengono descritti, ordinati e spazializzati, interconnessi in ambiente digitale, su livelli cartografici e testuali, attraverso un lavoro collaborativo e interdisciplinare, al fine di consentire l'analisi interpretativa integrata del problema storiografico con la sua lettura su scale analitiche diverse e interconnesse e la comparazione tra differenti studi di caso.

PAROLE CHIAVE

Storia digitale; storia del Mediterraneo; atlante storico digitale; Public history.

1. QUESTIONE DI RICERCA

La forma spaziale del regno di Sardegna era quella dell'arcipelago. Il territorio che ancora oggi chiamiamo Sardegna è in realtà formato da circa quaranta isole, le principali delle quali hanno generate esperienze storiche specifiche, irriducibili a quelle dell'isola maggiore sotto i profili linguistici e culturali e dunque storici.

Nonostante la forte relazione stabilita nel corso dei secoli tra questo regno e il Mediterraneo, la Sardegna è ancora oggi vista e concepita come una terra senza il mare. Lo è sotto il profilo storiografico dal momento che la grande parte degli studi storici sono relativi alle sue economie interne, il cui profilo non è stato mai letto sistematicamente in rapporto alle reti mediterranee in cui il regno si trovava giocoforza inserito.

Le analisi della dimensione politico-istituzionale della Sardegna medievale, condotte da vari studiosi come Antonio Era, Francesco Loddo Canepa e Bachisio Raimondo Motzo, hanno originariamente inserito la regione nel contesto più ampio del Mediterraneo catalano, come indicato anche da Alberto Boscolo. Nel corso del Novecento, questa prospettiva è stata integrata da una visione influenzata dalla Nouvelle Histoire francese, focalizzata sui quadri territoriali come spazi storici ed economici unitari, nonché sui processi socio-economici e fattori ambientali di lunga durata [8, 10, 18]. Questa analisi è stata anche arricchita da una prospettiva quantitativa [1].

A partire dagli anni Ottanta, sono emersi studi microstorici dedicati a comprendere più approfonditamente la società rurale isolana durante l'antico regime [14]. Un'attenzione particolare è stata rivolta ai rapporti di potere "informale", come i circuiti di corte, le reti clientelari e i bandos, che hanno giocato un ruolo significativo nell'integrazione del regno nella geopolitica mediterranea asburgica e sabauda [12].

Nel contesto di queste ricerche, Marco Tangheroni ha enfatizzato la dimensione marittima del regno di Sardegna, concentrandosi su quadri urbani con la presenza di mercanti stranieri [17]. Questo invito allo studio della mercatura marittima è stato accolto successivamente da diversi studiosi medievali [19]. Per l'età moderna, sono stati condotti studi approfonditi su colonie liguri e mercanti tra Bosa, Alghero e Cagliari, nonché indagini sul corallo, peschiere e saline. Ulteriori ricerche si sono concentrate sulle fortificazioni urbane e costiere sarde, in connessione con le attività economico-produttive lungo la costa. L'analisi complessa del rapporto del regno con il mare emerge nelle opere della storiografia internazionale che si è occupata della guerra.

Il rapporto del regno col mare emerge tra le righe della storiografia internazionale che si è occupata di guerra di corsa [3, 6, 11], commerci mediterranei, schiavitù [3], storia navale [11], contrabbandi [7]. Numerose sono le tracce disponibili nelle fonti archivistiche (da Parigi a Valencia, da Marsiglia a Genova, da Venezia a Livorno a Firenze, da Minorca, a Marsiglia, a Nizza e a Napoli).

Tuttavia, il quadro delle nostre conoscenze è frammentario, marcato da forti dislivelli sul piano degli approfondimenti monografici e, dunque, piuttosto incompleto. Il progetto di ricerca intitolato *The Digital Atlas of Maritime Sardinian history* (ASMSA) è nato per affrontare simile lacuna storiografica. La necessità di ordinare i materiali esistenti sulla questione e

di integrarli in un'unica piattaforma informativa con quelli generati dalla ricerca ha spinto a scegliere l'Atlante quale strumento più adatto allo scopo.
L'Atlante, nella sua concezione comune, rappresenta uno strumento di raccolta e pubblicazione di mappe geografiche, arricchite da testi e altri elementi informativi. Questo strumento assume una connotazione storica quando documenta eventi e fenomeni culturali legati a diversi periodi passati e a specifici territori. L'Atlante storico diventa digitale quando la sua capacità informativa e analitica si amplia o potenziata attraverso l'uso di applicativi, metodi e linguaggi computazionali.
A livello internazionale, il termine "Atlante storico digitale" è stato impiegato per descrivere varie iniziative che differiscono per struttura, obiettivi, metodi e dati utilizzati. A titolo di esempio, il *Digital Atlas of European Historiography since 1800* offre un quadro dell'impatto generato dai processi di istituzionalizzazione e professionalizzazione della storia sulle diverse storiografie nazionali attraverso una serie di mappe[1]. Allo stesso modo, il *Digital Atlas of Roman Sanctuaries in the Danubian Provinces (DAS)* utilizza Google Maps per mappare i santuari romani nelle province danubiane[2]. Infine, l'*Atlas of the Historical Geography of the United State*s rappresenta un'altra espressione di Atlante storico digitale[3]. L'*Atlas of the Historical Geography of the United States* è una versione digitale georeferenziata dell'opera omonima pubblicata da Charles O. Paullin e John K. Wright nel 1932. Altri esempi di Atlanti storici digitali includono l'*Irish Historic Towns Atlas* (IHTA), che utilizza il GIS per registrare lo sviluppo topografico di città irlandesi integrando dati testuali[4], e *Slave Voyages*, un progetto che ricostruisce le migrazioni forzate di schiavi africani negli spazi atlantici tramite banche dati integrate con mappe e linee temporali[5]. HGIS Germany pubblica mappe interattive per la storia territoriale tedesca, mentre DARE geolocalizza insediamenti romani su una cartografia digitale, documentandosi attraverso varie fonti accessibili direttamente sulla mappa[6].
Il progetto ASMSA prende ispirazione da queste esperienze, ma si sviluppa per rispondere a proprie esigenze di ricerca, affrontando la frammentarietà delle informazioni sulla storia marittima sarda per creare un quadro conoscitivo completo, integrando bibliografia di riferimento, documentazione estratta dagli archivi e materiali multimediali.
La banca dati digitale così costruita è immaginata non solo e non tanto quale strumento disseminativo, quanto soprattutto di lavoro a disposizione i ricercatori del progetto, per consentire loro di raccogliere autonomamente dati in modo condiviso e poi di interconnetterli e incrociarli per aumentare la capacità interpretativa della singola ricerca alla luce di quelle condotte dagli altri componenti dell'equipe.

2. GRUPPO DI RICERCA

Finanziato nel 2020 dalla Fondazione di Sardegna, ASMSA nasce dal lavoro di un gruppo interdisciplinare composto da esperti di diverse discipline, tra cui Storia moderna, archivistica, geografia, storia del pensiero politico, paleografia, archeologia. Il gruppo di lavoro ha preventivamente sviluppato un sistema di procedure e applicativi software tali da consentire di descrivere, ordinare, classificare, spazializzare e interconnettere dati e informazioni in ambiente digitale, dando forma a livelli cartografici, iconografici e testuali, per finalità sia interpretative che disseminative. Obiettivo era dotare ciascun ricercatore di uno strumento e un metodo che gli/le consentisse di raccogliere dati in maniera del tutto autonoma, ma secondo una regola condivisa, in modo da permettere lo scambio di informazioni tra ricerche condotte con approcci disciplinari diversi sugli stessi oggetti, o con i medesimi approcci ma su oggetti differenti; occorreva offrire un metodo di lavoro per acquisire, descrivere e ordinare qualsiasi tipologia di informazione, a prescindere dal supporto sul quale essa si trovava depositata; bisognava infine rendere la raccolta elettronica di informazioni un'esperienza semplice e accessibile a chiunque, anche e soprattutto a chi ha poca o nessuna familiarità col digitale.

3. STRUMENTAZIONE E FLUSSO DI LAVORO

Al centro dell'intervento AIUCD 2024 ci sarà il processo di definizione di uno strumento in grado di adattarsi agli sviluppi anche imprevisti della ricerca, nei diversi livelli di approfondimento in cui essa si articola è stata realizzata nell'ecosistema digitale (DOG) sviluppato nel DH UNICA - centro interdipartimentale di Umanistica Digitale - e grazie al supporto del LUDiCa (laboratorio di umanistica digitale dell'Università di Cagliari).

[1] Universität Trier, 2022, https://daeh.uni-trier.de/map/.
[2] 2018, https://danubianreligion.com/atlas-of-roman-sanctuaries-in-the-danubian-provinces/
[3] Andrew W. Mellon Foundation. *Atlas of the Historical Geography of the United States*. The Digital Scholarship Lab Un. of Richmond, https://dsl.richmond.edu/historicalatlas/.
[4] 1981, https://www.ria.ie/research-projects/irish-historic-towns-atlas
[5] https://www.slavevoyages.org/
[6] Università di Lund, 2015, https://imperium.ahlfeldt.se/

Obiettivo di questa prima fase progettuale è stato dare corpo a un flusso di lavoro collaborativo e interdisciplinare che consentisse l'analisi interpretativa integrata del problema storiografico con la sua lettura su scale analitiche e da prospettive diverse, con la comparazione tra differenti studi di caso.
In secondo luogo, l'analisi preliminare puntava a rendere l'intero processo accessibile anche a ricercatori non abituati a lavorare in ambiente digitale.
Sono stati elaborati e testati strumenti e procedure, definendo modelli di scheda descrittiva e schemi di lavorazione per la raccolta di informazioni documentali e multimediali necessarie alla strutturazione dell'Atlante.
Sotto il profilo infrastrutturale l'Atlante opera attraverso tre applicativi: Drupal, Omeka-S e Geonode. Si tratta di un ecosistema basato su tecnologie docker e git, strutturato secondo i principi della infrastructure as a code, che ha consentito finora di realizzare ambienti versionati e replicabili, attraverso l'impiego di applicativi quali Wordpress, Omeka Classic e dei già citati Omeka-S, Drupal, Geonode. Lo sviluppo di una serie di widget ha esteso la nativa funzione layout builder di Drupal, rendendo possibile comporre pagine web arricchite di informazioni presenti in un archivio Omeka, organizzate in gallerie e visionabili in forma geolocalizzata su mappe, sulle quali è possibile sovrapporre uno o più layers wms esterni, prodotti con Qgis e pubblicati con Geonode.
Ha preso così forma ASMSA, che è in grado di operare come mezzo di raccolta e ordinamento delle informazioni esistenti e come spazio di sintesi testuali e spaziali per l'analisi integrata e a diverse scale della dimensione marittima del regno, con la capacità di visualizzare l'intrecciarsi di azioni pubbliche e private, di processi generali con pratiche locali, dentro un quadro multi-livellare, diacronico e polifonico [16].

4. PROCESSARE LE FONTI

Vediamo più nel dettaglio il funzionamento di questo metodo di raccolta dati. Le informazioni estratte dalla bibliografia vengono metadatate con software Zotero e poi importate su Omeka-S. I documenti raccolti con gli scavi archivistici, le fotografie storiche o quelle effettuate dai ricercatori sul campo, le eventuali memorie orali depositate su formati audio-visivi, le cartografie storiche ecc., vengono invece descritte attraverso fogli di calcolo organizzati per colonna, ognuna delle quali corrispondente a un metadato della scheda descrittiva preventivamente definita dal gruppo di ricerca su Omeka-S. Esiste un foglio di calcolo/scheda descrittiva per ciascuna tipologia documentaria. I metadati di ciascuna schede descrittiva sono estratti dalle ontologie normalizzate dai rispettivi ambiti di ricerca.
Una volta terminata la compilazione dei fogli di calcolo, questi possono essere importati massivamente su Omeka-S e tradotti in schede descrittive pubblicabili online e geo-localizzate.
L'infrastruttura consente di processare anche oggetti generate anche da ricerche d'ambito non propriamente storico (nel nostro caso si pensi all'archeologia marina o all'antropologia impegnata sul mondo della pesca) e, qualora lo si ritenga necessario, con la partecipazione di diverse tipologie di pubblico.
I fogli di calcolo vengono impiegati anche per la costruzione di livelli di analisi spaziale attraverso il software Qgis. Questi livelli possono essere sovrapposti e combinati ad altri livelli generati georeferenziando cartografia storica o layer pubblicati da enti pubblici (nel caso di specie sono stati utilizzati i big data cartografici pubblicati nel Geoportale della Regione Autonoma della Sardegna). Attraverso Geonode i progetti costruiti con Qgis possono essere pubblicati e successivamente richiamati dentro l'ambiente Drupal dove è possibile geolocalizzarvi le schede prodotte con Omeka-S.
Il flusso di lavoro attivato grazie all'ecosistema digitale è sintetizzato nella figura 1.

5. RICERCA E PUBLIC ENGAGEMENT

Uno degli sviluppi potenziali più interessanti di ASMA è proprio quello relativo alla messa in campo di azioni di Public History e public engagement finalizzate a coinvolgere le comunità nella co-generazione di diverse tipologie di informazione: dalla micro-toponomastica custodita nella memoria locale, alle fotografie storiche conservate negli archivi familiari, passando per l'individuazione sul campo di strutture architettoniche o di pratiche legate alla cultura del mare nelle sue diverse manifestazioni, materiali e immateriali.
Il coinvolgimento del pubblico è importante poi per l'analisi dei processi di memorializzazione che investono temi e problemi marittimi. Si pensi, per fare un esempio, ai musei o ai culti dei santi locali variamente legati alla storia del mare, alle forme pubbliche di rievocazione di eventi o momenti della storia insediativa comunitaria. Si tratta di strumenti di studio e acquisizione dati già testati nell'ambiente del LUDiCa.
ASMSA si configura certamente come esperienza di Digital Humanities, ma aperta anche al campo vasto delle Public Humanities.

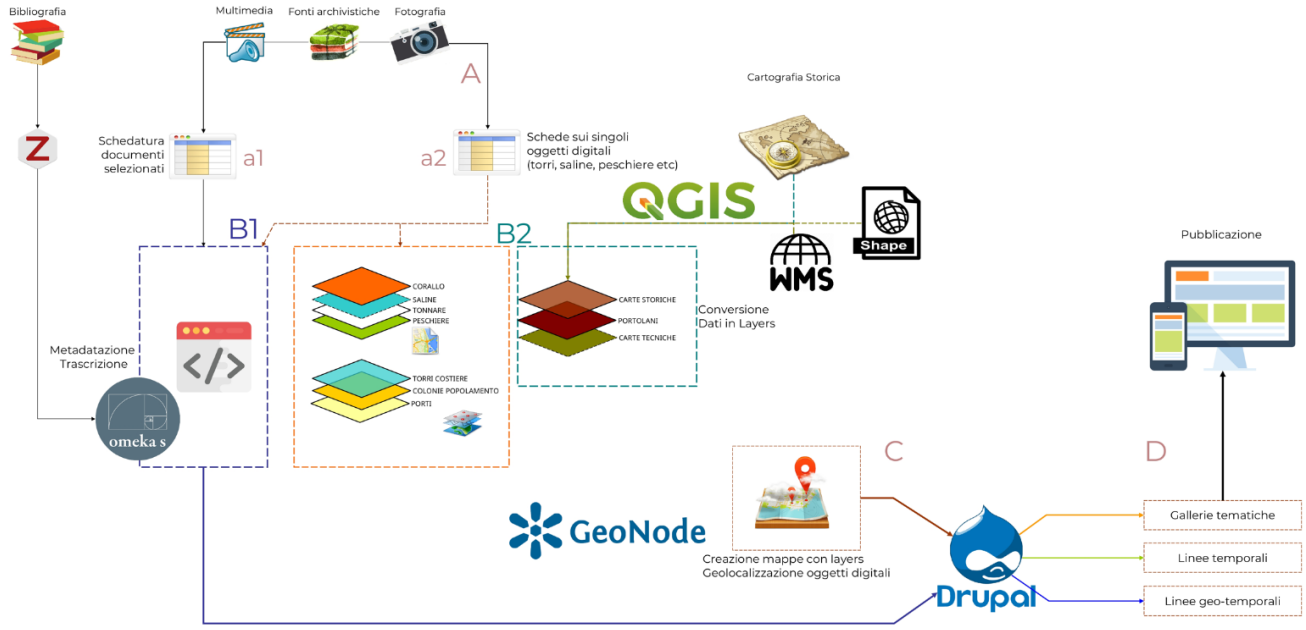


Figura 1. Flusso di lavoro in ASMSA

BIBLIOGRAFIA

[1] Anatra, Bruno. «Economia sarda e commercio mediterraneo nel basso Medioevo e nell'Età moderna». In *Storia dei sardi e della Sardegna. L'età moderna. Dagli aragonesi alla fine del dominio spagnolo*, a cura di Massimo Guidetti, III:109–216. Milano: Jaca Book, 1989.

[2] Basso, Luca Lo. *In traccia de' legni nemici: corsari europei nel Mediterraneo del Settecento*. Ventimiglia: Philobiblon, 2002.

[3] Bono, Salvatore. *Guerre corsare nel Mediterraneo: una storia di incursioni, arrembaggi, razzie*. Bologna: Il Mulino, 2019.

[4] Bono, Salvatore. *Schiavi: una storia mediterranea (XVI-XIX secolo)*. Bologna: Il Mulino, 2016.

[5] Boscolo, Alberto. «I Catalani nel Mediterraneo nel Basso Medioevo: aspetti e problemi». *Nuova rivista storica* 68 (1984): 1–20.

[6] Braudel, Fernand. *La Méditerranée et le monde méditeranéen à l'époque de Philippe II*. Vol. 3. Paris: Armand Colin, 1990.

[7] Calcagno, Paolo. *Fraudum: contrabbandi e illeciti doganali nel Mediterraneo (sec. XVIII)*. Roma: Carocci, 2019.

[8] Day, John. *Uomini e terre nella Sardegna coloniale: XII-XVIII secolo*. Torino: Celid, 1987.

[9] Doneddu, Giuseppe. «La pesca negli stagni di Oristano in età moderna». In *Giudicato d'Arborea e Marchesato di Oristano: proiezioni mediterranee e aspetti di storia locale. Atti del 1. Convegno internazionale di studi*, a cura di Giampaolo Mele, 1–2:487–506. Oristano: ISTAR, 2000.

[10] Le Lannou, Maurice. *Pâtres et paysans de la Sardaigne*. Tours: Arroult, 1941.

[11] Lo Basso, Luca. *Capitani, corsari e armatori. I mestieri e le culture del mare dalla tratta degli schiavi a Garibaldi*. Novi Ligure: Città del Silenzio, 2011.

[12] Manconi, Francesco. *La Sardegna al tempo degli Asburgo: secoli XVI-XVII*. Vol. 5. Nuoro: Il Maestrale, 2010.

[13] Mele, Giuseppe. «La rete commerciale ligure in Sardegna nella prima metà del XVII secolo». In *Génova y la monarquía hispánica (1528-1713)*, a cura di Manuel Herrero Sánchez, Yasmina Rocío Ben Yessef Garfia, Carlo Bitossi, e Dino Puncuh, 1:203–18. Società ligure di storia patria, 2011.

[14] Ortu, Gian Giacomo. *L'economia pastorale della Sardegna moderna: saggio di antropologia storica sulla «soccida»*. Cagliari: Edizioni Della Torre, 1981.

[15] Pira, Stefano. *Storia del commercio del sale tra Mediterraneo e Atlantico*. Cagliari: AM&D, 1997.

[16] Salice, Giampaolo. «ASMSA. Atlante digitale per la storia marittima della Sardegna». *Umanistica digitale*, fasc. 15 (2023): 117–32. https://doi.org/10.6092/issn.2532-8816/16660.

[17] Tangheroni, Marco. *Aspetti del commercio dei cereali nei Paesi della Corona d'Aragona. 1. La Sardegna*. Pisa: Pacini, 1981.

[18] Terrosu Asole, Angela. «La nascita di abitati in Sardegna dall'alto medioevo ai nostri giorni». In *Atlante della Sardegna*, di Roberto Pracchi, Vol. 2. Roma: La zattera, 1979.

[19] Tognetti, Sergio. «Il ruolo della Sardegna nel commercio mediterraneo del Quattrocento. Alcune considerazioni sulla base di fonti toscane». *Archivio Storico Italiano* 163, fasc. 1 (603) (2005): 87–132.

Un futuro per la memoria. Strumenti, modelli e sinergie per l'integrazione dei dati nel Portale delle fonti per la storia della Repubblica italiana

Michela Tardella[1], Roberta Maggi[2], Giorgia Lodi[3], Riccardo Albertoni[4], Herbert Natta[5], Gianluca Rossi[6], Tiziana Pasciuto[7], Paola Ciandrini[8], Luca Sinopoli[9], Maria Teresa Artese[10], Isabella Gagliardi[11], Eleonora Lattanzi[12], Sara Ventroni[13], Elisa Tizzoni[14], Alessandro Russo[15], Margherita Porena[16]

[1] CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italia – michela.tardella@cnr.it
[2] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – roberta.maggi@cnr.it
[3] CNR Istituto di scienze e tecnologie della cognizione, Italia – giorgia.lodi@istc.cnr.it
[4] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – riccardo.albertoni@ge.imati.cnr.it
[5] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – herbert.natta@ge.imati.cnr.it
[6] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – gianluca.rossi@ge.imati.cnr.it
[7] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – tiziana.pasciuto@ge.imati.cnr.it
[8] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – paola.ciandrini@ge.imati.cnr.it
[9] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – luca.sinopoli@ge.imati.cnr.it
[10] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – teresa@mi.imati.cnr.it
[11] CNR Istituto di Matematica applicate e tecnologie informatiche, Italia – gagliardi@mi.imati.cnr.it
[12] CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italia – eleonora.lattanzi@cnr.it
[13] CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italia – sara.ventroni@iliesi.cnr.it
[14] CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italia – elisa.tizzoni@gmail.com
[15] CNR Istituto di scienze e tecnologie della cognizione, Italia – alessandro.russo@cnr.it
[16] CNR Istituto di scienze e tecnologie della cognizione, Italia – margherita.porena@istc.cnr.it

ABSTRACT

Il *Portale delle fonti per la storia della Repubblica italiana* nasce dalla volontà di offrire a un pubblico ampio e diversificato una raccolta di fonti per la storia politica e istituzionale nazionale relativa alla seconda metà del Novecento. L'intento è creare un'architettura software e un'infrastruttura dati in grado di favorire l'accesso effettivo al patrimonio culturale considerato nel progetto; un accesso il più possibile aperto, che agevoli la conoscenza dei documenti conservati nei complessi archivistici degli organi costituzionali e dagli apparati amministrativi dello Stato, integrati con quelli delle associazioni private che concorrono alla vita democratica del paese, in particolare le fonti prodotte dai partiti politici e dalle organizzazioni sindacali. Una sfida culturale e tecnologica importante, ricca di implicazioni civili, ma anche teoriche e metodologiche, che vede impegnati tre istituti del CNR in collaborazione con soggetti pubblici e fondazioni e istituti privati.

PAROLE CHIAVE

Cultural heritage; Valorisation; Data Infrastructure; Linked Open Data.

1. INTRODUZIONE

Il *Portale delle fonti per la storia della Repubblica italiana* nasce dalla volontà di offrire a un pubblico ampio e diversificato una raccolta di fonti per la storia politica e istituzionale nazionale relativa alla seconda metà del Novecento. L'intento dei soggetti promotori dell'iniziativa è di creare un'architettura software e un'infrastruttura dati in grado di favorire l'accesso effettivo al patrimonio culturale oggetto del progetto; un accesso il più possibile aperto che agevoli la conoscenza dei documenti conservati nei complessi archivistici degli organi costituzionali e dagli apparati amministrativi dello Stato, integrati con quelli delle associazioni private che concorrono alla vita democratica del paese, in particolare le fonti prodotte dai partiti politici (art. 49 della Costituzione) e dalle organizzazioni sindacali (art. 39). Una sfida culturale e tecnologica importante, ricca di implicazioni civili, ma anche teoriche e metodologiche, che vede impegnati tre istituti del CNR in collaborazione con soggetti pubblici e fondazioni e istituti privati.

L'impegno teorico maggiore è stato dedicato all'elaborazione concettuale finalizzata a mettere a sistema risorse caratterizzate da una eterogeneità diffusa, per quanto riguarda sia le modalità e gli strumenti di descrizione delle risorse sia i formati e le tecnologie di trasmissione e condivisione dei dati. Sono ampiamente note, infatti, le specificità di formazione, gestione e conservazione di complessi archivistici prodotti dalle istituzioni pubbliche e private, nonché la vischiosità che caratterizza il rapporto esistente fra la vita istituzionale e la dimensione politico-sociale. Basta, ancora oggi, sfogliare la *Guida generale agli archivi di Stato*, curata da Claudio Pavone e Piero D'Angiolini o ancora il volume *Gli Archivi dei*

partiti politici (che raccoglie gli atti dei seminari tenuti a Roma e Perugia del giugno-ottobre 1994) per avere un quadro della complessità conservativa e descrittiva delle fonti a nostra disposizione [1, 8]. In questo orizzonte, il *Portale* ambisce a diventare uno spazio virtuale, un ecosistema nel quale arrivare ad una integrazione dei dati e dei sistemi, superando quel 'particolarismo informativo' che caratterizza l'eterogeneità e la frammentarietà delle informazioni, che si rispecchia anche nella proliferazione dei siti per la ricerca archivistica [3, 11][1]. Per ottemperare a queste finalità, il nostro lavoro si è posizionato nella dimensione della sinergia, intendendo con questo termine un'azione coordinata e contemporanea di più elementi, suscettibile di numerose declinazioni: sinergia, appunto, tra i patrimoni conservati dai diversi soggetti coinvolti; sinergia di modelli concettuali per la descrizione e la rappresentazione del patrimonio culturale; sinergia di linguaggi e formati di rappresentazione e condivisione dei contenuti e dei contesti[2]. Data questa specifica connotazione del progetto, è chiaramente cruciale la contaminazione e l'integrazione tra esperti di vari ambiti disciplinari (archivisti, bibliotecari, informatici, ma anche filosofi, esperti di teoria dei linguaggi), che lavorando in sinergica armonia, mirano ad una integrazione di vari domini in cui nessuna delle competenze risulta essere ancella dell'altra. Per gli scopi descritti è stato dunque concepito un sistema che mira ad una piena interoperabilità, progettato per ingerire, gestire ed armonizzare grandi quantità di dati in formati e modelli eterogenei.

2. PER UN'INTEGRAZIONE MODULARE

La progettazione e sviluppo dell'infrastruttura digitale del Portale è stata orientata da scelte tecnologiche volte a ottimizzare sia l'interoperabilità del sistema con fonti dati esterne sia l'interazione di utenti e agenti software con il livello di presentazione, declinato in diverse soluzioni di interfaccia.

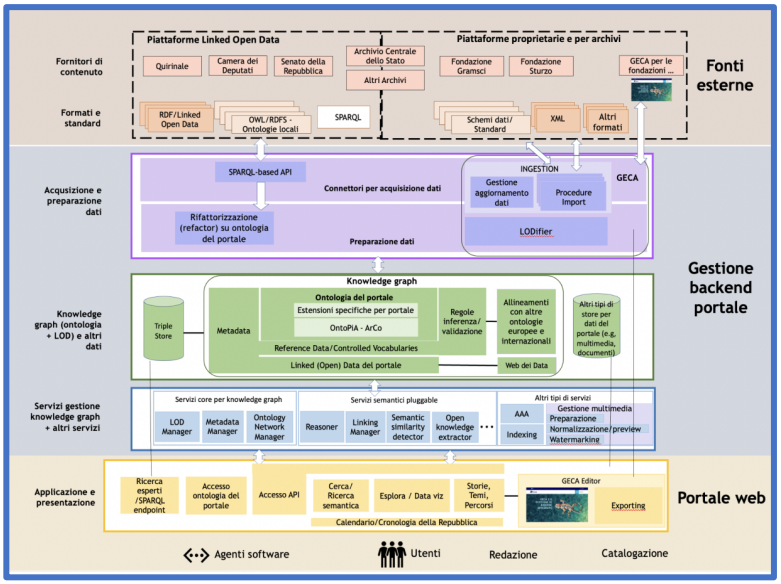


Figura 1. Schema architetturale del sistema con definizione dei livelli logici e delle componenti funzionali

Nella figura 1 è rappresentata l'architettura tecnica di alto livello del Portale, implementata sfruttando il modello architetturale a microservizi, che favorisce la possibilità di modificare e integrare modularmente le componenti open del sistema (la modularità, è utile precisare, caratterizza tutte le linee di attività del progetto e le metodologie utilizzate per il suo sviluppo, sia sul piano architetturale nel suo complesso, sia per la realizzazione delle ontologie). La parte centrale (in verde) costituisce la peculiarità di questa architettura che si struttura, oltre che intorno alle componenti software, anche attraverso il paradigma rappresentativo del knowledge graph. È opportuno soffermarsi sul doppio percorso di ingestione dei dati:

1. Per i portali istituzionali che dispongono di dati strutturati in formato Linked Open Data - ci riferiamo in

[1] Sebbene particolare riguardo sia riservato al patrimonio archivistico, l'infrastruttura dedicata è aperta anche all'acquisizione e gestione di risorse bibliografiche e museali, garantendo la possibilità di trattare in modo trasversale i vari domini del patrimonio culturale.

[2] L'eterogeneità dei dati condivisi, in diversi formati, che variano da documenti testuali (.doc) e tabellari (.csv, .xls) e a dataset in XML e JSON strutturati in base a modelli definiti dagli standard di dominio e/o dalle specifiche dei sistemi software che li hanno originati, ha richiesto la definizione di specifiche *pipeline* di estrazione e trasformazione, finalizzate all'integrazione nel modello logico-semantico del Portale, basato su framework RDF e sul modello dei Linked Open Data.

particolare al portale dell'Archivio Storico della Presidenza della Repubblica e a quello dell'Archivio Storico della Camera - si è proceduto con l'acquisizione dei dati esposti. Per quanto riguarda l'Archivio Storico del Senato, si è proceduto con la produzione di Linked Open Data (LOD) dei dati relativi alle persone, trasmessi al gruppo di lavoro in formato JSON e ricavati anche attraverso i LOD del Senato della Repubblica.

2. Per quanto riguarda invece i soggetti privati sono state predisposte le procedure di importazione di dati e di oggetti digitali, che hanno confermato la difformità delle situazioni che devono essere gestite: alcuni enti, infatti, hanno inviato i propri strumenti di ricerca in formato word, altri in excel accompagnato dal relativo pdf per fini di controllo, altri in XML (diversificati, in ragione della molteplicità dei in relazione alla diversità dei software di descrizione utilizzati o delle loro customizzazioni).

3. IL PERCORSO GECA

Le descrizioni e i documenti provenienti dalle istituzioni pubbliche e private che non dispongono di LOD (parte destra dello schema architetturale in Figura 1) vengono acquisiti attraverso il software GECA [6, 9]. Nato dall'evoluzione di GECA System (2004) e GECA RDC (2014) - strumenti per la catalogazione e la gestione di risorse bibliografiche - GECA è stato sviluppato come strumento destinato a differenti domini, quali il contesto museografico (tracciati ICCD, come schede F, OA, PST) e il contesto archivistico (complessi archivistici, unità archivistiche e documentarie, soggetti produttori e conservatori) proponendosi come collettore di descrizioni (cataloghi e strumenti di ricerca), nel rispetto dei relativi standard di settore (UNIMARC, MARC21, ISAD(G), EAD).

In GECA trovano piena attuazione tanto gli standard internazionali di dominio, quanto l'interoperabilità con i principali tracciati di interscambio, rappresentando quindi uno strumento utile sia alla granularità delle descrizioni previste dai diversi ambiti di applicazione, sia all'attivazione di relazioni cross dominio. Un esempio in questo senso è dato dalla possibilità fornita dal software di strutturare un albero archivistico (vd. Fig. 2) utilizzando non solo schede di descrizione per complessi archivistici e unità, ma anche schede dedicate a risorse bibliografiche o museali, integrate quindi con la descrizione archivistica [7].

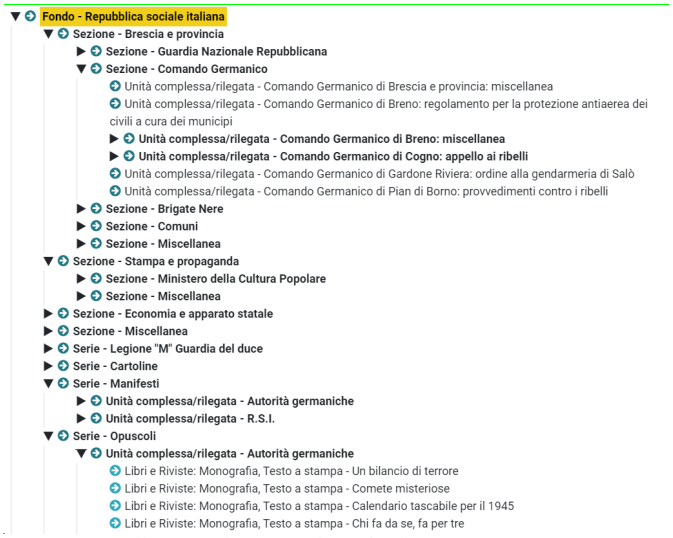


Figura 2. L'albero archivistico di un fondo inserito in Geca, contenente sia risorse archivistiche sia bibliografiche

L'integrazione di GECA come componente modulare all'interno dell'architettura del sistema ha quindi l'obiettivo, da un lato, di fornire uno strumento utile all'inserimento nativo di nuove descrizioni, utilizzando tracciati specifici per i tre domini di riferimento, e, dall'altro, di importare dati garantendo sia la compatibilità con standard riconosciuti (quali ICAR-IMPORT, EAD, ecc.) sia la flessibilità necessaria a gestire l'eterogeneità delle fonti e dei formati di interscambio (XML, JSON, CSV, ecc.).

GECA offre inoltre un modulo di lodificazione che permette di esporre le risorse in catalogo attraverso il paradigma dei LOD, utilizzando ontologie di dominio di riferimento internazionale (quali Bibframe e RiC-O) e tenendo in considerazione le esperienze nazionali (quali SAN Lod, OAD ed ArCo). Il processo di lodificazione di GECA (modulo LODifier) si distingue per l'adozione di standard consolidati a livello internazionale e software a licenza libera, garantendo così l'assenza di dipendenze da software proprietari. Un esempio tangibile è l'utilizzo dello standard internazionale W3C R2RML come linguaggio per la definizione delle mappature dal modello relazionale, nativo in GECA 3, agli insiemi di dati RDF. Per

l'archiviazione dell'RDF e l'interpretazione degli script R2RML, si fa affidamento sul software Open Link Virtuoso Universal Server, il quale offre una soluzione per gestire questa trasformazione e la gestione dei dati in formato RDF. Il processo di lodificazione si raccorda con la/le ontologie realizzate/usate appositamente per il Progetto, fornendo sia le specificità dei singoli domini, sia la trasversalità delle diverse risorse del portale e permettendo in questo modo non solo l'esposizione in LOD dei dati, ma anche la loro integrazione con altre fonti già lodificate.
Nell'ottica di ottimizzare l'integrazione cross dominio, e di massimizzare le potenzialità offerte dalle tecnologie semantiche impiegate, è in corso di revisione il modello dati relativo alle voci di autorità, che rappresentano entità fondamentali per attivare relazioni all'interno delle risorse in catalogo. In particolare, la revisione riguarda principalmente la modellazione di un'entità agente, che rappresenta in modo univoco enti, persone, famiglie, tipizzata in funzione del tipo di relazione con le risorse (soggetto produttore, conservatore, autore, ecc.), e di altre entità (quali luoghi, soggetti) funzionali a esplicitare sia le connessioni interne al catalogo sia esterne (tramite rimandi a sistemi di rappresentazione quali VIAF, Geonames, Wikidata), massimizzando le possibilità offerte dai LOD in termini di arricchimento della conoscenza e recupero del contenuto informativo.
Tra le altre funzionalità offerte dal software, di particolare rilevanza per il Progetto è la possibilità di creare percorsi di esplorazione delle risorse su base tematica: una caratteristica che ha permesso di estendere l'ambito applicativo del modulo GECA dalla gestione del catalogo alla gestione dei contenuti previsti per il portale, permettendo di inserire e modificare prodotti redazionali e contenuti audio/video e di creare relazioni tra questi e le altre tipologie di risorse presenti nel sistema. La modularità e flessibilità di GECA ne permette l'utilizzo anche in contesti applicativi diversi dall'architettura del portale, rappresentando uno strumento utile sia per la creazione e gestione di basi dati di beni culturali, sia per la loro pubblicazione ed esplorazione.

4. IL GRAFO DELLA CONOSCENZA DEL PORTALE

La parte sinistra dell'architettura di Figura 1 illustra gli strumenti e i servizi destinati all'acquisizione dei dati disponibili secondo il paradigma dei Linked Open Data, ovvero rappresentati attraverso standard del Web Semantico quali RDF. Da un punto di vista metodologico, si è avviata una prima analisi dei modelli dei dati (o ontologie) definite e usate dalle fonti nella pubblicazione dei loro LOD, con un particolare focus sulle fonti Camera dei Deputati, Senato della Repubblica, Presidenza della Repubblica e Archivio Centrale dello Stato. Dall'analisi è emerso un chiaro requisito di armonizzazione dei modelli dati esistenti, utile per il raggiungimento degli scopi specifici del portale. Sono poi stati raccolti requisiti mediante un'interazione diretta con esperti di dominio, facendo uso dello strumento delle domande di competenza (chiamate in inglese Competency question) ovvero domande che gli utenti potrebbero porre sui dati di loro interesse. Una buona modellazione infatti deve poter supportare interrogazioni, anche complesse, sui dati che siano capaci di rispondere a tali domande. Infine, si è svolta un'analisi dello stato dell'arte rispetto a ontologie esistenti a livello nazionale, più o meno vincolanti anche nel contesto dell'ecosistema di interoperabilità nazionale (nel caso degli archivi: SAN LOD[3] e OAD[4], nel caso delle persone CPV[5] pubblicata nel catalogo nazionale della semantica schema.gov.it, Cultural-ON - Cultural Institute or Site ONtology [5]), rispetto a standard internazionali di riferimento (e.g., RiC-O per il settore archivistico, FRBR per il contesto bibliografico/lavori) e, infine, rispetto a cosiddetti ontology design pattern che potessero essere riutilizzati con successo rispetto alle esigenze di modellazione del portale (e.g., il pattern del ruolo che cambia nel tempo per agenti). Gli ontology design pattern (ODP) sono soluzioni riusabili usate per risolvere problemi ricorrenti nella modellazione ontologica. È stato provato che il riutilizzo di ODP migliora la qualità dell'ontologia, riduce l'arbitrarietà nel design dell'ontologia e favorisce così l'interoperabilità semantica dei dati [4, 10]. Tutte queste fasi rientrano nel contesto più ampio della metodologia eXtreme Design (XD) [9, 10] particolarmente efficace per lo sviluppo agile di ontologie o modelli dati. Dai requisiti raccolti è risultato tuttavia chiaro che quanto esistente poteva essere limitato per rispondere ad alcune esigenze del portale. Pertanto, si è proceduto con la progettazione di una vera e propria rete di moduli ontologici dove un approccio modulare è stato scelto in quanto particolarmente efficace in presenza di tipologie diverse di dati con, potenzialmente, esigenze di aggiornamento diverse. La modularizzazione infatti agevola la manutenzione complessiva in ambito software e di modellazione ontologica. La rete di ontologie sviluppate è mostrata in Figura 2. In particolare, sono stati creati otto moduli ontologici[6] che rappresentano le persone e le figure politiche, i loro incarichi, i loro mandati, le organizzazioni costituzionali, istituzionali e politiche, gli eventi pubblici di particolare rilevanza pubblica come l'elezione del presidente della repubblica, i bollettini dei vari rami del parlamento, i discorsi, gli atti e i diari storici nonché le principali

[3] https://www.san.beniculturali.it/web/san/dati-san-lod
[4] https://labs.regesta.com/progettoReload/oad-ontology/
[5] https://w3id.org/italia/onto/CPV
[6] Tutte le ontologie prodotte sono disponibili come risorse aperte su piattaforma github al seguente link https://github.com/PortaleFontiRepubblica/assets/tree/main/ontologies

informazioni sugli archivi (modulo in Figura 2), non più rappresentati come solo informazioni di catalogazione ma come beni culturali con una loro identità. Questa particolare modellazione degli archivi, differente da modelli esistenti allo stato dell'arte è stata possibile grazie all'integrazione del modulo archivi nella rete di ontologie nazionali sui beni culturali (ArCo) [2] che porta appunto una prospettiva diversa di modellazione dei beni: il focus non è più la scheda di catalogo ma il bene inteso come entità culturali con una propria identità. A questi moduli ontologici della rete se ne aggiunge uno di servizio chiamato IRHO (Italian Republic History Ontology) che consente l'accesso a tutti i moduli ontologici in un unico punto in quanto importa tutti gli altri (nodo centrale della rete di Figura 3).

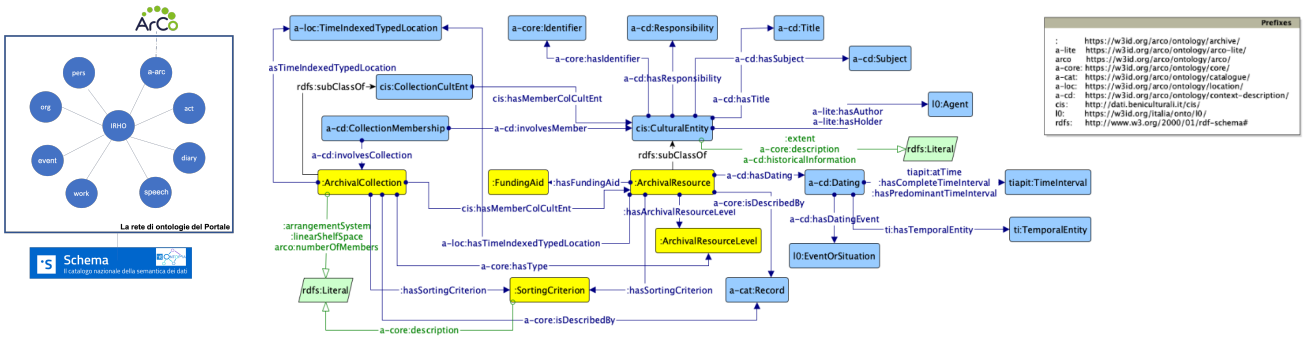


Figura 3. La rete di ontologie del portale e il focus sul modulo degli archivi integrato in ArCo

5. COMUNICARE IL PATRIMONIO

Il tema, sempre attuale, della comunicazione/divulgazione del patrimonio culturale passa attraverso l'analisi e l'applicazione di molteplici aspetti, quali l'uso di metodologie e tecnologie innovative, ma anche la messa in atto di strumenti di approfondimento orientati ad ampliare la fruizione da parte di un pubblico diversificato. Consapevoli dell'importanza civile di coinvolgere tipologie di utenti dai profili eterogenei, si è agito su un duplice livello: le modalità di accesso agevolato ai documenti della storia repubblicana (mantenendo il legame con le informazioni di contesto); l'ideazione di prodotti di natura divulgativa appositamente pensati per questa iniziativa progettuale. A tal fine, è stato costituito un coordinamento di storici contemporaneisti, che lavora in sinergia con gli altri esperti di dominio e con i membri della redazione. È stato elaborato un piano editoriale articolato, la cui ossatura è costituita da approfondimenti relativi ai principi costituzionali e da 14 temi correlati. In tale ottica, sono stati realizzati approfondimenti testuali relativi alla genesi dei primi 12 articoli della Costituzione e una serie di videointerviste e podcast che discutono e contestualizzano le risorse documentali: la centralità delle fonti, in quanto oggetto di studio e di preservazione, trova piena attuazione anche in questa linea di attività. Per la gestione delle tre tipologie di approfondimento è stata sviluppata una scheda di metadatazione (vd. Fig. 4), la cui messa a punto ha tenuto conto proprio dell'esigenza di metterle in relazione con le fonti che ne costituiscono l'oggetto. Anche questa linea di attività contribuisce quindi ad aumentare le possibili interrelazioni e integrazioni tra beni diversi, nell'ottica propria dell'apertura e della trasversalità.

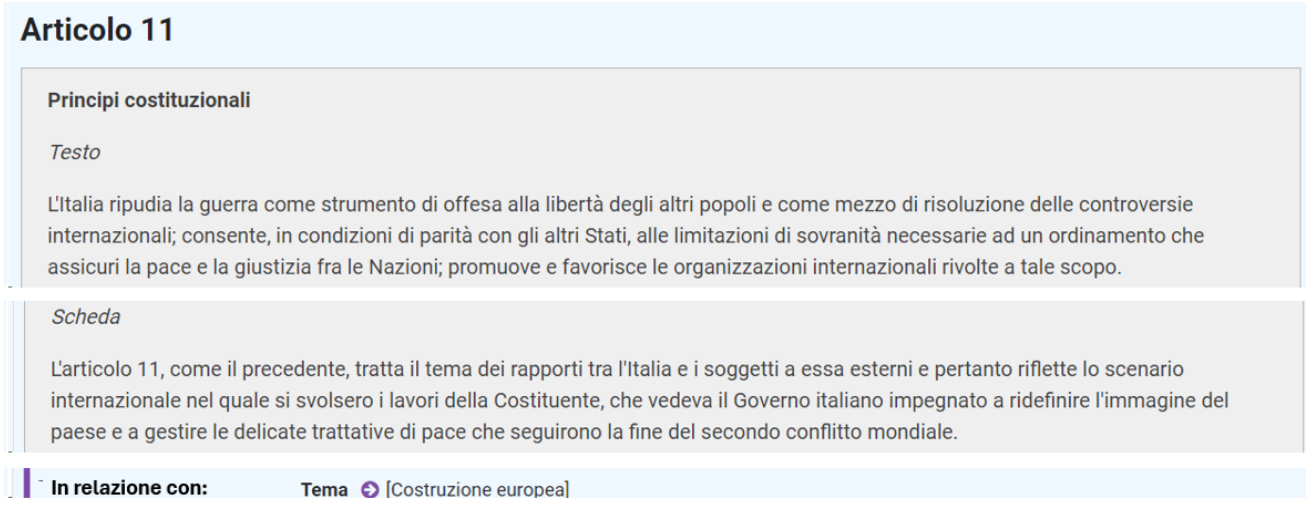


Figura 4. Anteprima di una scheda di metadatazione di un prodotto redazionale, personalizzata rispetto ai contenuti previsti per il Portale delle Fonti

Un ulteriore obiettivo del progetto consiste nella creazione di uno spazio dedicato alle scuole, con la realizzazione di percorsi didattici, strutturati a partire da documenti d'archivio, utilizzabili dagli insegnanti nel corso delle loro lezioni. La costruzione dei percorsi, realizzati attraverso le metodologie e le tecnologie riferibili alla didattica della storia, è supervisionata dal coordinamento dei consulenti storici.

BIBLIOGRAFIA

[1] Cacioli, Manuela. *Gli archivi dei partiti politici: atti dei seminari di Roma, 30 giugno 1994, e di Perugia, 25- 26 ottobre 1994*. Roma: Ministero per i beni culturali e ambientali, Ufficio centrale per i beni archivistici, 1996.

[2] Carriero, Valentina A., Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, e Chiara Veninata. «ArCo: the Italian Cultural Heritage Knowledge Graph». In *The Semantic Web - {ISWC} 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part {II}}*, a cura di Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, e Fabien Gandon, 36–52. Lecture Notes in Computer Science. Auckland, New Zealand, 2019. https://dx.doi.org/10.1007/978-3-030-30796-7_3.

[3] Ciandrini, Paola, Eleonora Lattanzi, Roberta Maggi, e Michela Tardella. *Archivi e contaminazioni disciplinari: dai linguaggi ai modelli, dagli approcci alle tecniche*. Plurimi. Cnr-edizioni, in corso di stampa. https://www.cnr.it/it/plurimi.

[4] Gangemi, Aldo, e Valentina Presutti. «Ontology design patterns». In *Handbook on ontologies*, a cura di Steffen Staab e Rudi Studer, 221–43. Berlin: Springer, 2009.

[5] Lodi, Giorgia, Luigi Asprino, Andrea Giovanni Nuzzolese, Valentina Presutti, Aldo Gangemi, Diego Reforgiato Recupero, Chiara Veninata, e Annarita Orsini. «Semantic web for cultural heritage valorisation». In *Data Analytics in Digital Humanities*, 3–37. Springer, 2017.

[6] Maggi, Roberta, Tiziana Pasciuto, Martina Mazzoleni, Maria Teresa Artese, Isabella Gagliardi, e Riccardo Albertoni. «GECA 3.0 - A new tool for cataloguing and enjoying cultural heritage». *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 373-379, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[7] Pasciuto, Tiziana, Riccardo Albertoni, Roberta Maggi, Maria Teresa Artese, Isabella Gagliardi, e Maurizio Gentilini. «Travelling Culture: Define, Implement, Enrich and Disseminate the Digital Cultural Heritage. The "DigitXL Project" Case Study». In *EDEN Research Workshop, Dubrovnik, 19-20/09/2022. Proceedings*, a cura di Josep M. Duart e Elena Trepule, 134–39. Dubrovnik, 2022.

[8] Pavone, Claudio, e Piero D'Angiolini. *Guida generale degli archivi di Stato italiani*. Vol. 1–4. Roma: Ministero per i beni culturali e ambientali, Ufficio centrale per i beni archivistici, 1981.

[9] Presutti, Valentina, Enrico Daga, Aldo Gangemi, e Eva Blomqvist. «eXtreme Design with Content Ontology Design Patterns». In *Proceedings of Workshop on Ontology Patterns*, a cura di Eva Blomqvist, Kurt Sandkuhl, Scharffe, Francois, e Svatek, Vojtek. CEUR Workshop Proceedings. CEUR-WS.org, 2009.

[10] Presutti, Valentina, Giorgia Lodi, Andrea G. Nuzzolese, Aldo Gangemi, Silvio Peroni, e Luigi Aprino. «The role of ontology design patterns in linked data projects». In *Proceedings of Conceptual Modeling: 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016*, a cura di Isabelle Comyn-Wattiau, Katsumi Tanaka, Il-Yeol Song, Shuichiro Yamamoto, e Motoshi Saeki, Vol. 9974. LNCS. Springer, 2016. https://doi.org/10.1007/978-3-319-46397-1_9.

[11] Zanni Rosiello, Isabella. «Sul mestiere dell'archivista». In *L'archivistica sul confine. Scritti di Isabella Zanni Rosiello*, 371–88. Roma: Poligrafico Zecca dello Stato, 2000.

INFRASTRUTTURE PER LA RICERCA UMANISTICA

Cophi Editor: From GreekSchools to the projects multiverse

Simone Zenzaro
CNR Istituto di Linguistica Computazionale, Italy - simone.zenzaro@cnr.it

ABSTRACT

CoPhi Editor is a web platform designed for digital scholarly editing, specifically developed to assist papyrologists in creating a digital critical edition of Philodemus' Arrangement of the Philosophers. While initially tailored for the unique challenges of digital papyrology, CoPhi Editor has the potential for broader adoption. To facilitate its use across various domains, we undertook a process of generalization of the tool implementing the concept of *project*. In this paper, we outline how this generalization process has been conducted.

KEYWORDS

Domain Specific Languages; Computational Philology; Digital Philology; DSE tools.

1. INTRODUCTION

CoPhi Editor [6, 7] is a web-based collaborative and cooperative authoring platform for Digital Scholarly Editions (DSE) that is being developed as part of the ERC AdG 885222-GreekSchools project. Among the main goals of CoPhi Editor there is the will to provide the scholars with an authoring text-centered platform with minimal learning curve. To achieve this goal we adopted the DSL-based DSE methodology [2] that enables scholars to stick to their well established editorial conventions while allowing the automatic generation of a TEI/EpiDoc compliant digital edition from the edited texts.

Cooperation and collaboration are the other two directions in which CoPhi Editor is moving. The editing platform is designed to implement cooperation through a collaborative review process that involves threads of discussions based on annotations of the texts and the consequent dialogue among scholars leading to the *constitutio textus*. Such annotation are compliant with the Web Annotation Data Model (WADM)[1]. A text in CoPhi Editor is also editable concurrently by different editors preserving the text integrity à la Google Docs. This feature enables collaboration between users.

From a technical standpoint, the platform's architecture (see Fig. 1) is built upon microservices, with each one dedicated to a specific isolated feature. These features include authentication, authorization, data storage, IIIF[2] viewer, DSL management, collaboration among users, and the web application itself, serving as the users' entry point for interacting with texts and witnesses.

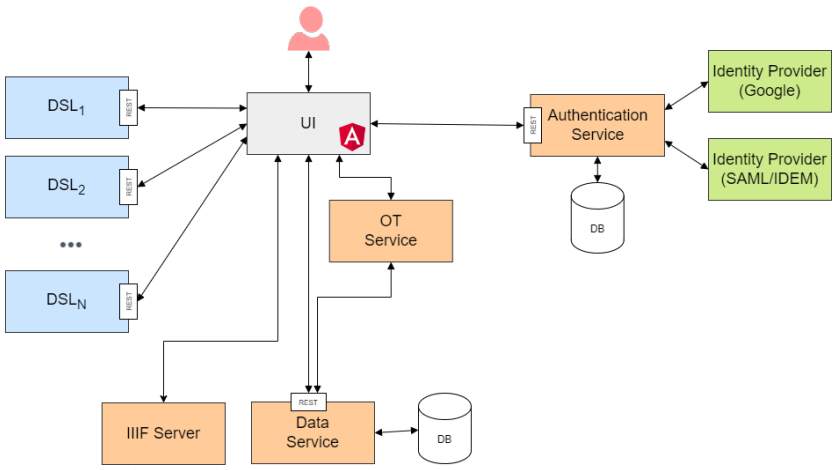


Figure 1. CoPhi Editor Architecture

CoPhi Editor was initially developed to support the editorial challenges of digital papyrology. As the web platform can be generalized to domains beyond its original scope, its design was oriented towards transforming it into a domain-agnostic and cross-domain environment.

[1] https://www.w3.org/TR/annotation-model/
[2] https://iiif.io/

2. MAKING COPHI EDITOR DOMAIN-AGNOSTIC

In order to make CoPhi Editor a domain agnostic-platform, the first activity is to systematically identify the GreekSchools specific features within the platform and then try to generalise them. In particular we have identified three major points to be generalised: the set of Domain Specific Languages (DSLs) [4, 5] that define the texts' domain, the notion of *project* in the application, and the access restrictions to the application based on per project user roles.

The set of DSLs for a given context in the application is directly related to the *types of texts* that the scholar wants to encode. A typical example of DSL in CoPhi Editor is the apparatus. In the GreekSchools project, for example, there are five different DSLs that represent the *diplomatic* and *literary* transcriptions, *paleographic* and *philological* apparatuses, and the Italian *translation* for a given column of papyrus. For each DSL, the CoPhi Editor web application interacts, via a shared RESTful [3] Application Programming Interface (API), with a web service that primarily offers the capability to request syntactical error checking for editorial conventions and provides contextual suggestions for text completion. In this regard, the first step toward a domain agnostic platform has been fulfilled by the choice to model the architecture as microservices. As a result, within the DSL-based environment of CoPhi Editor, achieving support for multiple domains is possible by providing the corresponding DSL service with the only constraint to comply with the shared REST API definition inspired by the Language Server Protocol[3]. Anticipating the potential expansion of DSLs for use in CoPhi Editor, we are planning to introduce a *registry* service that acts as a proxy between the available DSL services and the actual web application itself. The registry service should collect, filter, and provide the DSLs while the web application should select the core subset of DSLs for the project.

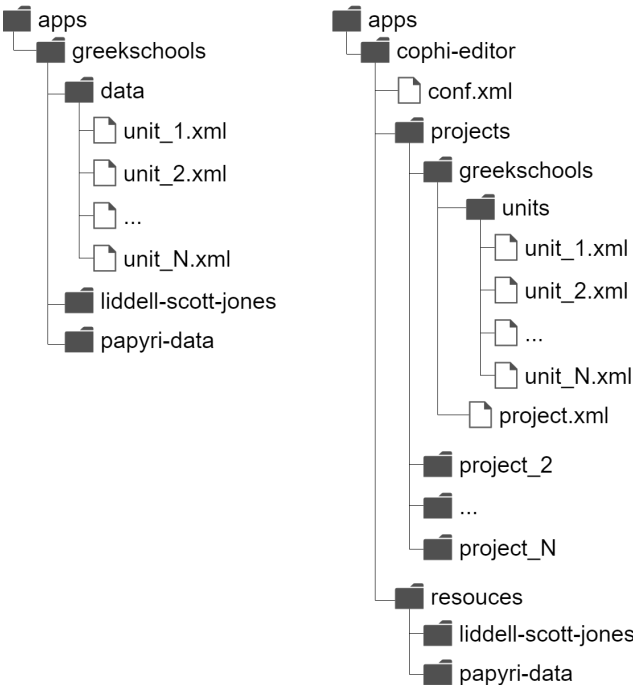


Figure 2. Database folder restructuring

The second aspect to tackle involves supporting the concept of *projects*, which can be further subdivided into three distinct subtasks: project configuration, database structure refactoring, and API changes. A CoPhi Editor *project* is the set of information that plays a role in the creation of a Digital Scholarly Edition. In particular a *project* contains a set of metadata and data. The metadata describe a title, the owner of the project, a set of team members, a set of roles, the set of resources linked to the project, and the set of DSLs references. The data is represented by the set of DSL texts that are called *units* within CoPhi Editor. A unit is a recursive data structure based on [1] that represents any portion of the text from entire works, single chapters, annotations or even smaller segments of text. Projects should be self contained to avoid interference between different projects. To keep them organised, each project is described by a configuration file that, in our current technological stack, is defined by an XML file inside exist-db[4] that contains all the above mentioned information. To actually avoid contaminations between projects, we needed to change the database structure. Figure 2 shows on the left the original folder structure and on the right the new structure. Notably, the original structure is flat and contains an application

[3] https://microsoft.github.io/language-server-protocol/
[4] https://exist-db.org/

named *greekschools* where the units are stored inside the data folder. Every other material, for example the Liddell-Scott-Jones Greek-English Lexicon (liddell-scott-jones) or other sources of papyrological texts (papyri-data) is available next to the data folder. The new folder structure instead has a single root folder (cophi-editor) that contains a global application configuration (conf.xml) and the two folders *projects* and *resources*. The *resources* folder is the place where every project can collect additional resources (e.g. searching the lexicon) and is kept separated from the project itself to reduce redundancy in case multiple projects use the same resources by referencing all the needed resources in their specific project configuration. The *projects* folder contains the list of active projects in CoPhi Editor. Each project folder has a configuration file (project.xml) and the *units* folder that corresponds to the original *data* folder. Mixing this database structure and the access control via roles it is possible to protect the access to the project data and support multiple projects inside the same web application.

Likewise the database structure, also the *data service* API needed some changes. The *data service* is a web service with a RESTful API that is in charge of accessing and persisting the units' data. The REST endpoints were unaware of the project to which the units belong. In order to avoid any disruptive changes we decided to add a single parameter in the header section of the API endpoints. This way the project info does not pollute the API paths keeping them concise yet understandable.

Finally, Cophi Editor supports a project agnostic definition of roles. Each role describes which data a user can access or modify. A role definition is based on the endpoints of the data service API organised in *activities.* So a role is a set of activities. An activity corresponds to a granted operation on a resource possibly restricted by some parameters of the actual API. Accessing a resource involves identifying an activity within the user's roles that grants the necessary access.

Although the *data service* is the same across different projects, the definition of distinct roles for each project results in diverse behaviors concerning resource access restrictions.

Combining all of these changes, CoPhi Editor is now ready to host multiple projects.

3. ACKNOWLEDGEMENTS

The GreekSchools project, "The Greek Philosophical School according to Europe's earliest 'history of philosophy': Towards a new pioneering critical edition of Philodemus' Arrangement of the Philosophers" has received funding from the European Research Council (ERC) under the European Union's Horizon 2020, Excellent Science (Grant agreement No. 885222, PI: Graziano Ranocchia).

REFERENCES

[1] Del Grosso, Angelo Mario, D. Albanesi, E. Giovannetti, and S. Marchi. 'Defining the Core Entities of an Environment for Textual Processing in Literary Computing»'. *DH2016*, 2016, 771–75.

[2] Del Grosso, Angelo Mario, Simone Zenzaro, Federico Boschetti, and Graziano Ranocchia. 'GreekSchools: Making Traditional Papyrology Machine Actionable through Domain-Driven Design'. In *IEEE CiSt'23 – 7th IEEE Congress on Information Science & Technology*, 2023.

[3] Fielding, Roy Thomas. 'REST: Architectural Styles and the Design of Network-Based Software Architectures'. Doctoral dissertation, University of California, 2000.

[4] Fowler, Martin. *Domain-Specific Languages*. Londra: Pearson Education, 2010.

[5] Parr, Terence. *Language Implementation Patterns Create Your Own Domain-Specific and General Programming Languages*. Pragmatic Bookshelf, 2014.

[6] Zenzaro, Simone, Angelo Mario Del Grosso, Federico Boschetti, and Graziano Ranocchia. 'Ease the Collaboration Making Scholarly Editions: The GreekSchools Case Study'. In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 230-232, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[7] Zenzaro, Simone, Angelo Mario Del Grosso, Federico Boschetti, and Graziano Ranocchia. 'Verso la definizione di criteri per valutare soluzioni di scholarly editing digitale: il caso d'uso Greekschools'. In *AIUCD 2022 - Culture Digitali. Intersezioni: Filosofia, Arti, Media. Proceedings Della 11a Conferenza Nazionale, Lecce, 2022*, edited by Fabio Ciracì, Giulia Miglietta, and Carola Gatto, 20–25. Quaderni di Umanistica Digitale, 2022. https://doi.org/10.6092/unibo/amsacta/6848.

Digital Humanities and Heritage Science: moving from landscaping to a dynamic research observatory in an Open Science Cloud

Roberta Bianca Luzietti[1], Alessia Spadi[2], Nicola Giampietro[3], Giacomo Mancuso[4], Alessandra Caravale[5], Antonio D'Eredità[6], Marta Caradonna[7], Paola Moscati[8], Valeria Quochi[9], Monica Monachini[10], Emiliano Degl'Innocenti[11]
[1]CNR Istituto di Linguistica Computazionale, University of Pisa, Italy, roberta.luzietti@phd.unipi.it
[2]CNR Istituto Opera del Vocabolario Italiano, Italy, alessia.spadi@cnr.it
[3]CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italy, nicola.giampietro@cnr.it
[4]CNR Istituto di Scienze del Patrimonio Culturale, Italy, giacomo.mancuso@cnr.it
[5]CNR Istituto di Scienze del Patrimonio Culturale, Italy, alessandra.caravale@cnr.it
[6]CNR Istituto di Scienze del Patrimonio Culturale, Italy, antonio.deredita@cnr.it
[7]CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italy, marta.caradonna@cnr.it
[8]CNR Istituto di Scienze del Patrimonio Culturale, Italy, paola.moscati@cnr.it
[9]CNR Istituto di Linguistica Computazionale, Italy, valeria.quochi@cnr.it,
[10]CNR Istituto di Linguistica Computazionale, Italy, monica.monachini@cnr.it
[11]CNR Istituto Opera del Vocabolario Italiano, Italy, emiliano.deglinnocenti@cnr.it

ABSTRACT
The contribution presents work, carried out in the second work package of the Humanities and Heritage Italian Science Cloud (H2IOSC) infrastructural project dedicated to 'Landscaping and Building Communities', on the definition of a methodology for landscaping the actual status of resource and technology availability and exploitation in the Humanities and Cultural Heritage. The activity involves a comprehensive investigation encompassing language technologies, digital humanities, and heritage science disciplines in Italy. The aim of the landscaping is to collect information on the latest and most prevalent resources, tools, communities, best practices, standards, and projects developed within the Heritage, Social Sciences, and Digital Humanities communities. In this work package, the four partnering infrastructures - CLARIN, DARIAH, E-RIHS, and OPERAS - collaborate closely to develop the best strategies for engaging and meeting the needs of their target research communities as well as to identify the set of priority items (resources, tools, and services) to FAIRify and onboard into the national Marketplace.

KEYWORDS
Research infrastructures; Digital Humanities; Heritage Science.

1. INTRODUCTION
This paper presents the collaborative work carried out in the second work package of the Humanities and Heritage Italian Science Cloud[1] (H2IOSC) infrastructural project where teams of the four Italian nodes of ESFRI[2], CLARIN[3], DARIAH[4], E-RIHS[5] and OPERAS[6] research infrastructures (RIs), aim at conducting a comprehensive investigation into the panorama of Italian digital resources within the digital humanities, linguistics, and heritage science disciplines.
A dedicated Landscaping Research Group (LRG), composed by representatives of the four RI, worked on collecting information concerning existing projects, digital resources, tools, communities, best practices, and standards in use among the research communities linked to the four RIs. Starting from an initial evaluation of the status of each RI, this mapping aimed at: i) identifying gaps in terms of new and existing data resources and technologies (i.e., tools, software, and services) not yet available through the RIs repositories and catalogues; ii) assessing their degree of FAIRness; and iii) complete a selection of the items to include into the RIs repositories and project Marketplace.
The rationale for this work is threefold: i) strengthen the competitiveness of RIs as valuable means for content search and deposit of data, tools, and services; ii) better meet and interpret the needs of different scholars and research communities;

[1] https://www.h2iosc.cnr.it
[2] https://www.esfri.eu
[3] www.clarin.eu, www.clarin-it.it
[4] www.dariah.eu, www.dariah.cnr.it
[5] www.e-rihs.eu, www.e-rihs.it
[6] https://operas-eu.org/

and iii) encourage the integration of new resources and tools into the project national cluster. The ultimate ambition of this landscaping initiative is to support RIs in the process of increasing the reliability that they can have among scholars and align with the evolving demands of the digital humanities, linguistics, and heritage science research communities. The combination of the different landscaping activities and tools will lead to the establishment of a permanent observatory to i) monitor the status of RIs in terms of new resources, best practices, technological and user needs and, at the same time; ii) contribute to the sustainability of the project over time.
In this contribution, section 2 explains the landscaping methodology; section 3 describes the information retrieval and annotation protocol; section 4 illustrates the design and dissemination of the questionnaire; section 5 presents the focus groups implementation plan; section 6 portrays the first version of the knowledge base elaborated for this landscaping task; finally, section 7, includes a brief discussion on the preliminary results and future project activities.

2. LANDSCAPING METHODOLOGY

To accomplish the landscaping task, a combination of qualitative and quantitative approaches, defined as Mixed Methods (see Fig. 1), is adopted[7] to:

- elaborate a picture of the current panorama and needs, as expressed by the research communities,
- support the activities of the other H2IOSC project working groups,
- support further analysis and forecasts related to user needs,
- support the elaboration of a long-term strategy for the implementation and development of the project observatory.

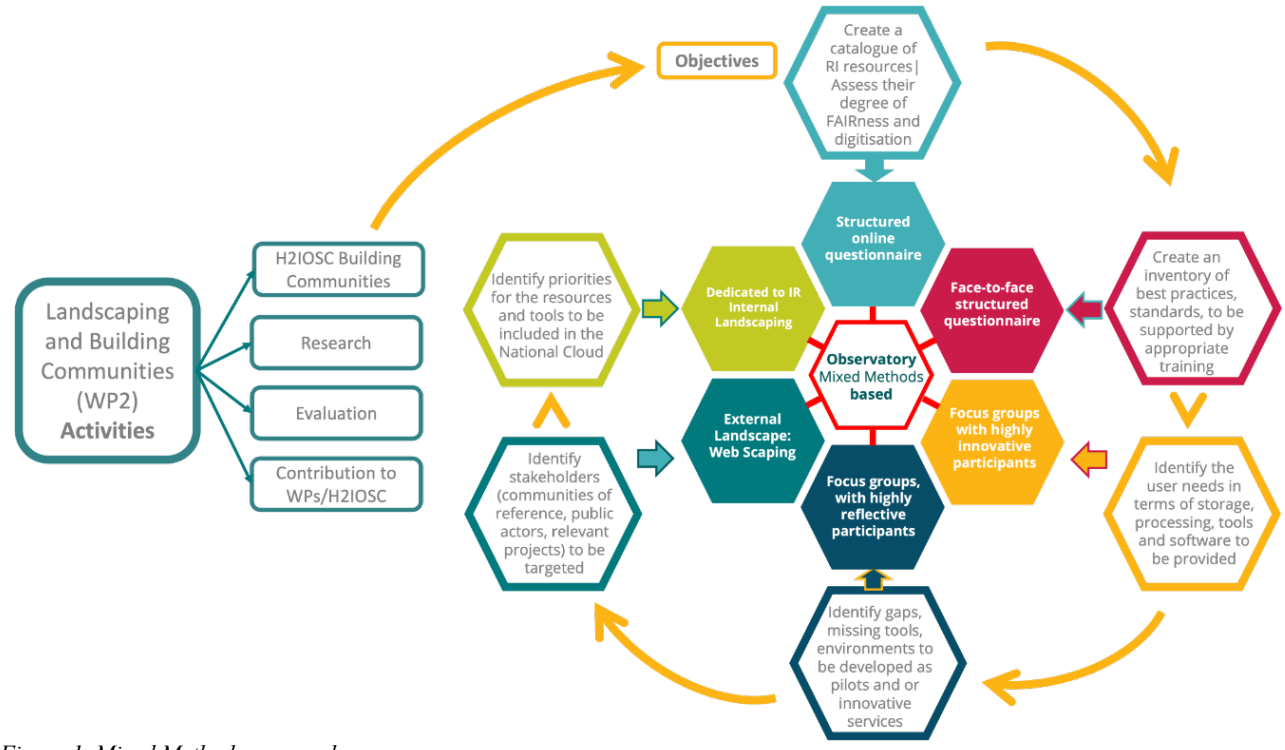


Figure 1. Mixed Methods approaches.

To design an apparatus capable of accounting for the existing projects, resources, tools, communities, best practices, and standards, in relation to each RI community involved in the project, we started from adopting and developing quantitative tools: i) a matrix to catalogue all the information retrieved encompassing projects, publications and conference proceedings in different disciplines, together with existing online catalogues and repositories; ii) a questionnaire to collect information directly from the research communities involved; and iii) a database to store and navigate all the acquired data. Strategic starting points for this investigation are the existing repositories and catalogs pertaining to two RIs involved in this project, which served as valuable initial sources of information (e.g. ILC4CLARIN[8]). At the same time, qualitative approaches, such as focus groups, are included to corroborate the quantitative results, and acquire new information on first impressions, critical aspects and needs of the target communities involved, as well as to stimulate interest in project activities. To support

[7] Among the various definitions of mixed methods, the one proposed by [6] best suits this case.
[8] https://ilc4clarin.ilc.cnr.it/, https://dspace-clarin-it.ilc.cnr.it/

the improvement, development, and deployment of the elaborated surveying tools, additional community building-oriented activities are put into plan, to attract and positively engage with the main social science, linguistics and humanities associations and representatives.
In practice, three main instruments have been developed to achieve the landscaping objectives: semi-automatic information gathering, an online survey, and focus groups. These instruments are detailed in the following sections and are designed to work in complementary ways to outline the landscape. The expectation is that cross analyzing the results from these instruments will provide valuable insights for research infrastructure (RI) managers and the H2IOSC working groups on several fronts:

- **prioritization**: identifying which data resources, tools, and services are more salient and need urgent integration into the RIs and the project Marketplace,
- **enhancements, FAIRification, and servification:** identifying resources that need to be improved and/or FAIRified,
- **gaps identification**: detecting the absence of crucial resources, tools, and services in existing RIs, and highlighting specific needs characterizing to the Italian research community,
- **training needs:** uncovering gaps in knowledge, competencies, and skills among community members for developing targeted training programs, materials, and campaigns.

3. INFORMATION GATHERING AND CLASSIFICATION

The creation of an observatory aimed at landscaping the panorama represented by the H2IOSC community in Italy - encompassing existing projects, resources, tools, communities, best practices, standards used, and needs - is a challenging task: the first steps towards this goal rely on an approach at the crossroads between the Mixed Methods methodology and the questionnaire. Moreover, it should be noted that the nature of the four participating RIs is fundamentally different, since they deal with different issues in many respects, including but not limited to data types/formats (e.g.: oral, written etc.) and disciplines/contents (e.g.: philology, cultural heritage, open science, linguistics, etc.). Therefore, the identification and classification of all the necessary information does not just require a shared strategy, but also an overall consideration of all the RIs specific features and needs. The goal here consists of mapping tools, datasets and projects supported by the RIs, emphasizing the importance of not just gathering data, but also gaining an overall understanding of the activities taking place within the different discipline domains. To effectively map resources across the four infrastructures, the definition of common practices and shared parameters is imperative. The methodology employed thus focuses on a detailed consideration of the current state of available resources and technological services within linguistic, humanistic, and heritage sectors of the research institutions involved. The heterogeneous range of mapped resources poses a challenge in defining a shared set of work tools to classify, group and correctly describe different resource types, standards, reference ontologies, vocabularies, taxonomies, and specific domain features. Particular attention is paid to the FAIRness assessment process involving the findability (online or in physical locations), accessibility, interoperability, and reusability in other contexts.
As the result of a collaborative effort among the four RIs, three data structures have been defined to map datasets, tools, and projects respectively. To describe each resource, the following pieces of information are thus recorded: i) general information about the resource; ii) data lifecycle management policies; iii) status (i.e., new, updated etc.). More in detail, among the parameters of interest: acronym, name, description, classification, context of use, curator, data format (for input and output), standards, licenses, notes are also recorded. In this work, extensive documentation is also developed to support the data entry process, so that each contributor can proceed easily. The tables gather a compilation of information to support other activities, like FAIRness evaluation of the resources. It is important to mention that criteria for recording information about resources and technologies established during this work are consistent in the design of the questionnaire, of the focus groups, and of the database, that will host all landscaping data to be made available to the permanent observatory. This will ensure that all the three instruments are coherent and interoperable.

4. QUESTIONNAIRE

The core activity of this workflow consists in conceiving a comprehensive questionnaire to directly collect information from the research communities regarding the following key aspects:

- usage and needs related to data resources and technologies,
- awareness and current exploitation of existing RIs services,
- desiderata for new services/offerings,

- training needs, as perceived by community members in terms of competences and skills that emerged throughout the survey, which is crucial for designing effective training programs, materials, and campaigns, especially at this initial stage.[9]

The decision to create a single questionnaire applicable to all RI communities of respondents was made to prevent data dispersion and mitigate the risk of "community overload" [5]. In fact, many researchers' fields of specialization are interdisciplinary, and, therefore, can be represented by more than one RI.

The work started from listing/pointing out all the expected information to be collected via an online [3]. Such outcomes consist in i) identification of stakeholders; ii) user needs; iii) training needs; iv) priorities; v) gaps, missing tools, and services to be developed; vi) existing resources (available through or outside the RIs); vii) degree of FAIRness, digitization of new resources and tools; and viii) best practices and standards. Subsequently, series of questions were formulated to acquire the necessary information for each objective (e.g., to know if a resource indicated by a respondent is FAIR compliant, we included four additional quick questions asking to indicate where the resource is deposited, catalogued, the PID and license). For each question, additional clarifications and information were included in the notes, to avoid as much as possible any sort of potential misunderstandings. For potentially equivocal or unfamiliar acronyms and terminology, such as FAIR[10] or PID[11], definitions and references were included to help participants better understand the questions and respond correctly. Such additional information might also generate curiosity on the topic. Lastly, each question was assigned to a type of response (open-ended, single-choice, or multiple-choice). Thus, the first final version of the questionnaire consisted of the following sections:

1. **Personal Information & Privacy Policy**: aimed to identify stakeholders; it contains questions about age, career level, research field, and type of institution/department of affiliation. All these common data were directly collected from the data subjects, who have been provided with a written informative prepared in accordance with the art. 13 of Regulation EU 2016/679 "General Data Protection Regulation" and treated anonymously and aggregated. Only e-mail addresses were employed for contact purposes only and upon expressed and informed consent.
2. **Data Resources, Software, Tools, and Technologies**: this section comprised four subsections to gather information about existing or newly created data resources, tools, etc. The goals here were multiple, from acquiring information on the state-of-the-art material within the Italian panorama, users' knowledge, expectations, and needs, as well as identifying gaps that should be addressed by other project's activities (e.g., 3 and 7).
3. **Projects**: to collect information on projects that develop(ed) data resources, digital tools, software, or other technologies relevant to respondents' disciplines (e.g., linguistics, archaeology, philology, etc.).
4. **Training needs**: a dedicated section was included to directly collect expectations and comments from participants, regarding their personal training needs.
5. **Prior knowledge of RIs**: aimed at assessing the extent of respondents' awareness and current usage of infrastructures, it is held to be crucial for planning informative and training actions to increase participation.
6. **Publications**: directed to both individual researchers and their institutions, it regards the publication of scientific articles. The questions here aimed to understand how individual researchers publish their research, whether they are aware of the publication practices adopted by their institution, and to what extent there is commitment for publishing in open access.
7. **Feedback**: aimed at collecting first impressions and insights on how respondents came across the questionnaire.

The questionnaire was designed to be distributed online, via Lime Survey[12]. The survey was developed and deployed stepwise in different successive stages. A first version was sent to a selected and restricted group of colleagues, representatives of different subcommunities, that acted as a test group and provided useful feedback that led to initial substantial improvements to the questionnaire. Subsequently, for statistical sampling purposes, the revised version of the questionnaire was forwarded to a control group consisting in few selected mailing-lists of relevant disciplinary associations for which we were able to know the exact numerosity. For the third and final version, currently in progress, the aim is to extend the investigation to a wider sample of respondents and to as many target community members as possible, using different dissemination channels such as: social networks, project and RIs websites, conference presentations, and dissemination events.

From a first analysis of the results obtained from the control group and the discussion with the representatives of the scientific communities, a revision of the questionnaire was necessary to make a better impact on the research community

[9] To do so, the LRG worked in strong collaboration with the project team dealing with training, also very involved with the target community.
[10] https://www.go-fair.org/fair-principles/
[11] https://www.clarin.eu/content/persistent-identifiers
[12] https://www.limesurvey.org/

of respondents and, at the same time, limit the possibility of receiving ambiguous answers. Some important adjustments aimed at: i) reducing the number of questions, by trying to merge and integrate them as much as possible, but without losing sight of the objectives of the project and the task; ii) supplementing the questions with more information, for facilitating and clarifying any doubts respondents might have about the type of answer to be given; and iii) allowing users to choose how to participate to the questionnaire. This last point emerged directly from the responses and opinions of researchers, university professors and head of research laboratories that found difficulties in answering exhaustively to questions in which they were asked to indicate, for example, the top ten resources created. Such needs gave rise to the proposal to divide the questionnaire into two sections. The first dedicated only to assess the type of respondents, regarding their experience with creation and/or usage of resources, knowledge of RIs and publication practices. The second, is optional and dedicated to describing the top five created or used resources. For the second questionnaire, respondents will be given the choice of filling out their responses online or via a face-to-face interview.

5. FOCUS GROUPS

The conduction of the focus groups will involve the participation of sixteen people, divided into two groups of eight people, balanced by gender. One group will be attended by researchers from both Italian universities and public research institutions (PRI), whereas the other group will be attended by master's and doctoral students.

The conduction of focus groups has several objectives:

- contribute to the construction of the project community,
- corroborate the survey results with qualitative data,
- gather qualitative data on the Digital Humanities experiences of the different audiences involved in the project,
- systematize and generalize data, after careful content analysis using the MAXQDA 2022 software[13].

From a methodological point of view, the focus groups will follow a tried and tested pattern in the social sciences [4:39-66; 1: 26-68] for such innovative field of inquiry in the discussion stimuli. Participants in the two groups will be invited to discuss their expectations on the outcomes, future developments, and long-term sustainability of the project. More specifically, researchers and students will be asked to i) make predictions about the kind and quantity of resources, services, and tools that they will have access to through the H2IOSC Cloud and ii) discuss about how, at a concrete level, the working group in charge of developing the marketplace is supposed to carry out these predictions [5: 13-22].

The motivation for integrating qualitative interviews is because, compared to other methodological constructs adopted so far, focus groups represent a more flexible instrument in terms of adaptability and reusability for future follow-ups.

Based on the definition in [2][14], these interviews will help the landscaping team gain an in-depth understanding of the preferences of the stakeholders in terms of resources, tools, and services related to all the communities involved in the project. Finally, focus group meetings will be conducted following the questionnaire format, while also leaving room for additional comments, ideas, and suggestions - providing valuable food for thought for the Marketplace implementation and the empowerment of the national RIs nodes, as well as for the elaboration of a long-term sustainability strategy for the H2IOSC RI federation.

6. DATABASES AND WEBAPPS

The decision to develop a database to store and analyze landscaping data comes from the acknowledgement of several needs within the H2IOSC research community, emerged during the works of the LRG, such as i) the necessity of creating a database for existing digital products, tools/software and research projects utilized in heritage science and digital cultural heritage; ii) the requirement to classify and index those products; iii) the need for analytical tools to help RIs adjust their strategic plans over time. Its primary objective is thus to gain a deeper understanding of the evolving landscape of digital cultural heritage/heritage science and to effectively address its requirements in terms of digital tools. Due to its purpose, it was named Digital Heritage Landscaping platfOrm (DHeLO). One of the first challenges was to structure a data schema able to fully represent the complexity and heterogeneity of the available data, while also allowing for sufficient segmentation for analytical purposes. Therefore, so far, five distinct entities have been identified (people, institutions, products, tools, and research projects), all interconnected with each other with multiple relations. This conceptual scheme has proven to be an efficient mapping solution for digital products developed for Social Sciences and Humanities, providing

[13] VERBI Software. (2021). MAXQDA 2022 [macOS Sonoma 14.4.1]. Berlin, Germany: VERBI Software. Available from https://maxqda.com/.

[14] We can define a (qualitative) interview as a conversation (a) provoked by the interviewer, (b) addressed to subjects selected based on a survey plan and (c) in substantial numbers, (d) having cognitive purposes, (e) guided by the interviewer, (f) based on a flexible, non-standardized question scheme [2: 401].

an in-depth insight into the current progress of the disciplines. The data collection process, ongoing, currently involves gathering information from relevant literature, incorporating well-known products and research projects, and the questionnaire results. Additionally, data from significant repositories used by the scientific community (e.g. Zenodo) are included. Currently the dataset holds more than 500 records from different sources, relevant for cultural heritage. DHeLO also enables seamless interconnection and interoperability with other platforms, facilitated by the release of an API in JSON format; it also supports statistical and spatial analysis using third-party software. Even at this early stage, the ongoing analysis of the collected data is progressively unveiling common practices, shared standards and similar workflows across various research projects and institutions. This holistic approach contributes to a more comprehensive understanding of the digital cultural heritage environment, fostering better cooperation and informed decision-making.

In parallel, the need also arose to develop a tool to systematically gather the rich sector-specific literature, the lifeblood of any research field, that plays a crucial role in the knowledge production process and the information dissemination system. In this perspective, it was decided to implement a large work of collection and systematization of these resources, named Bibliography of Digital Archaeology (BiDiAr), starting with the references published at the end of each article in the journal *Archeologia e Calcolatori*[15], a peer reviewed open access publication that has been a key scientific benchmark in the wide field of digital archaeology for over thirty years. The references were collected within Zotero, a well-renowned open-source software, commonly used by the scientific community. Trough Zotero API BiDiAr relates to DHelo, providing bibliographic references to the entry of the database. The cataloguing work has started with the volumes published in the last five years, with the idea to cover, over time, all the issues published since 1990. The volume of data collected (to date, more than 5000 entries) will provide an impressive bibliographic corpus related to the discipline, which will then be analyzed using various keys and tools. Analyses will highlight how ICT (Information and Communications Technologies) tools and techniques for studying the past have evolved over time, identify the currently most advanced research areas and the schools and teams most active in this research field, which appears to be constantly evolving and updating given the pervasive nature of technology in all its aspects.

7. CONCLUSION

This contribution gives a description of the methodology and initial activities implemented by the H2IOSC project for landscaping the panorama of resources, tools, services, and needs characterizing the digital humanities, linguistics, and heritage science disciplines. The information we aim to collect throughout this investigation is meant to help:

- better defining the project target research communities,
- laying the foundation for the creation of the project Marketplace,
- assigning the priorities for the integration of resources tools and services in the project Marketplace,
- providing information for the FAIRification of the identified resources to other project activities,
- creating a list of all the community training needs to be addressed by the dedicated project activity.

To do so we used a Mixed Methods approach. Starting from the above-mentioned objectives, we first planned the overall methodology defining relevant actors, information sources and landscaping instruments to be implemented. As a second step we proceeded in parallel by collecting data for the cataloguing matrix and by engaging directly with the RIs target communities, employing a two-way strategy: a questionnaire and two focus groups. After collecting the first results, we focused on constructing a database to reconcile all collected data and allow for smart visualization and interpretation.

The initial design of the questions in the questionnaire was further discussed in a meeting where we directly invited the RI communities to participate. Their feedback guided us towards the following adjustments: i) reducing the length of the questionnaire, without losing sight of the objectives of the project and the task; ii) supplementing the questions with more information to facilitate and clarify any doubts respondents might have about the answer to be given; and iii) allow the community to be more involved in the landscaping and research activities. Overall, the first key findings were that community participation to the questionnaire was unbalanced across the 4 RIs and, in general, 68.4% of the respondents are rather users than creators of new resources and tools. The most used and created resources are databases (semantic, relational, analytical), corpora (textual, oral), and 2D/3D models. Unfortunately, many of the resources mentioned by respondents are not yet openly available to the scientific community because they are not deposited online, accessible and/or easily findable. Furthermore, awareness of the FAIR principles and their impact on science appears inadequate; the same applies to the use of existing digital resources already available from the RIs certified repositories, an interesting insight for training activities. Nonetheless, one positive aspect is that, although the community did not seem to use RIs as

[15] https://www.archcalc.cnr.it

a first source of insight into resources, tools, and services, claimed to be very interested in learning more about the issues addressed by the project and in keeping up to date with its future developments.

8. ACKNOWLEDGEMENTS

This work is supported by the H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

REFERENCES

[1] Bezzi, Claudio. *Fare ricerca con i gruppi. Guida all’utilizzo di focus group, brainstorming, Delphi e altre tecniche.* Milano: Franco Angeli, 2020.

[2] Corbetta, Piergiorgio. *La Ricerca sociale: metodologia e tecniche: I paradigmi di riferimento*. Vol. I. Bologna: Il Mulino, 2014.

[3] Dewaele, Jean-Marc. ‘Online Questionnaires’. In *The Palgrave Handbook of Applied Linguistics Research Methodology*, edited by Aek Phakiti, Peter De Costa, Luke Plonsky, and Sue Starfield, 269–86. Palgrave Macmillan London, 2018.

[4] Frisina, Annalisa. *Focus Group, Una Guida Pratica*. Bologna: Il Mulino, 2010.

[5] Nardi, Peter M. *Doing survey research: A guide to quantitative methods*. New York: Routledge, 2018. https://doi.org/10.4324/9781315172231.

[6] Tashakkori, Abbas, and John W. Creswell. ‘The New Era of Mixed Methods’. *Journal of Mixed Methods Research* 1 (2007): 3–7.

Funzioni e sostenibilità di una piattaforma digitale per le lingue arcaiche

Michele Mallia[1], Riccardo Del Gratta[2], Valeria Quochi[3]
[1]CNR Istituto di Linguistica Computazionale “A. Zampolli", Italia - michele.mallia@ilc.cnr.it
[2] CNR Istituto di Linguistica Computazionale “A. Zampolli", Italia - riccardo.delgratta@ilc.cnr.it
[3] CNR Istituto di Linguistica Computazionale “A. Zampolli", Italia - valeria.quochi@ilc.cnr.it

ABSTRACT
Questo contributo, basato sull'esperienza acquisita in un progetto di ricerca triennale dedicato allo sviluppo di tecnologie e risorse digitali nel campo degli studi linguistico-storici su lingue epigrafiche frammentarie, riflette sulla sostenibilità a lungo termine dei risultati digitali ottenuti da piccoli gruppi di ricerca nelle Digital Humanities (DH). In particolare, l'analisi si concentra sulla possibilità di integrare questi risultati nelle infrastrutture di ricerca virtuali, distribuite e federate, come CLARIN(-IT) e la recente iniziativa di federazione delle infrastrutture di ricerca italiane per le Scienze Umane e il Patrimonio Culturale, denominata 'Humanities and Heritage Italian Open Cloud' (H2IOSC).

PAROLE CHIAVE
Epigrafia Digitale; Lessicografia Computazionale; OntoLex Lemon.

1. INTRODUZIONE
Negli anni recenti, la sostenibilità è emersa come un concetto cruciale e onnicomprensivo che influenza vari ambiti, dalla salvaguardia ambientale all'uso sostenibile di dati e strumenti nella ricerca scientifica e umanistica. In particolare, si evidenzia l'importanza della "FAIRificazione" dei dati e l'adozione di modelli, formati e metodologie che aderiscono a questo paradigma. Tra questi, il paradigma dei linked open data si sta diffondendo anche in ambito umanistico e storico, offrendo formati e modelli essenziali per la condivisione dei dati. Diversi studi, come quelli indicati in [2] e [3], confermano i numerosi vantaggi nell’adottare questi approcci, soprattutto nel campo delle digital humanities. Questi approcci mirano a garantire l’accessibilità e il riuso dei dati in maniera etica e trasparente, e favoriscono l’interconnessione tra le varie risorse, facilitando così la creazione di un ecosistema di conoscenze più ampio.
Nel settore specifico dell'epigrafia digitale, si assiste a una rapida evoluzione delle tecnologie con lo sviluppo di metodi, protocolli e progetti digitali per trattare anche le lingue poco documentate, storiche e arcaiche, come quelle trattate in [11]. Un esempio significativo è il progetto EAGLE[1] (Electronic Archive of Greek and Latin Epigraphy), che unisce diverse banche dati epigrafiche e stabilisce linee guida per produrre contenuti digitali standardizzati, favorendo l'interoperabilità e il riutilizzo dei dati. Altri progetti rilevanti e più recenti includono i.Sicily[2] [8], un corpus digitale di iscrizioni su pietra dal VII secolo a.C. al VII secolo d.C., e il corpus Cretan Institutional Inscriptions[3] [12], focalizzato sulle iscrizioni istituzionali di Creta. Entrambi adottano il modello TEI-EpiDoc per le iscrizioni, uno standard XML per codificare le iscrizioni, ma non integrano risorse lessico-concettuali e non mettono a disposizione dello studioso tecnologie integrate per la creazione o modifica di dati interconnessi. Il progetto LiLa (Linking Latin) [7], invece, rappresenta un'evoluzione tecnologica, integrando risorse linguistiche latine, sia lessico-concettuali sia testuali, secondo i principi FAIR e Linked Open Data, promuovendo l'interoperabilità e facilitando la connessione di risorse esterne. A differenza di iSicily o LiLa, molti altri progetti sono di breve-media durata e/o hanno scarso supporto infrastrutturale, e questo pone rischi di sostenibilità a lungo termine per i loro prodotti scientifici (specialmente software) in termini di manutenzione, distribuzione, e consultabilità pubblica.
In questo contributo, utilizziamo l'esperienza acquisita nel progetto PRIN “Lingue e Culture dell’Italia Antica. Linguistica Storica e modelli digitali” (ItAnt) per riflettere sulla sostenibilità a lungo termine dei risultati tecnologici ottenuti da piccoli gruppi di ricerca nel contesto di progetti di breve-media durata nel campo delle Digital Humanities (DH). Attraverso l'analisi della piattaforma DigItAnt, questo lavoro mira a delineare un quadro chiaro delle potenzialità e delle sfide legate all'integrazione delle tecnologie digitali negli studi antichistici, evidenziando come tali strumenti possano arricchire la comprensione e l'interpretazione dei dati storico-linguistici. Infine, esaminiamo in particolare le modalità con cui questi risultati possono essere mantenuti accessibili grazie alle infrastrutture di ricerca virtuali, distribuite e federate, come

[1] https://www.eagle-network.eu/
[2] https://sicily.classics.ox.ac.uk/
[3] https://ilc4clarin.ilc.cnr.it/cretaninscriptions/en/

CLARIN e la recente iniziativa Humanities and Heritage Italian Open Cloud (H2IOSC), la quale mira a federare quattro infrastrutture italiane in un cluster dedicato alle DH.
L'articolo è così strutturato: nella prima parte descriviamo la piattaforma DigItAnt, dedicata alla creazione e interrogazione di un insieme di dati interconnessi: lessici collegati a edizioni critiche di iscrizioni arcaiche, riferimenti bibliografici, vocabolari controllati). Ci concentreremo in particolare sull'applicazione di fruizione, un componente che non è stato ancora descritto.
Nella seconda parte riflettiamo sulla sostenibilità della piattaforma in ottica di migrazione verso l'infrastruttura di ricerca CLARIN-IT, al fine di garantire la sua disponibilità a lungo termine e offrire così un servizio affidabile alla comunità scientifica.

2. LA PIATTAFORMA DIGITANT

La piattaforma descritta in questo lavoro offre agli studiosi uno strumento avanzato per lo studio delle lingue arcaiche dell'Italia Antica[4], quali Osco, Falisco, Venetico, Celtico ecc., basato sulle loro evidenze testuali, consentendo loro di creare ecosistemi di dati collegati e di renderli accessibili e interrogabili a un pubblico più ampio. Il presente lavoro si inserisce nel contesto delle ricerche condotte sullo sviluppo di strumenti grafici interattivi per la creazione e l'interconnessione di dataset linguistici. In particolare, esso prende ispirazione da alcune esperienze precedenti, quali ad esempio EFES[5], che consente di creare interfacce per la consultazione di dati codificati in TEI-EpiDoc, ed estende le funzionalità di altri lavori, come ad esempio [1], per permettere di codificare lessici direttamente compatibili con il Web Semantico. Come illustrato in [9], questa applicazione consente di integrare voci lessicali con informazioni codificate in risorse indipendenti, in particolare edizioni di iscrizioni in TEI-EpiDoc, e di stabilire collegamenti azionabili con altri dataset esterni. L'applicazione mira anche a rendere questi dati fruibili e interrogabili attraverso un'interfaccia grafica intuitiva, arricchendo l'esperienza utente con un approccio che rispetta al contempo le prassi condivise delle discipline informatico-umanistiche e dei Linguistic Linked Open Data.
La progettazione della piattaforma DigItAnt è centrata su un'architettura orientata ai servizi (SOA) in cui le funzionalità di elaborazione, potenzialmente generiche, sono separate dalle interfacce grafiche rivolte all'utente e dunque specifiche per il caso d'uso particolare. DigItAnt espone due interfacce utente, corrispondenti a due modalità di utilizzo: un'applicazione è dedicata alla creazione e/o revisione di lessici storici (conformi al modello OntoLex-Lemon) collegati ai testi epigrafici (in XML secondo TEI-EpiDoc), alla bibliografia (gestita con Zotero) e ad altre risorse rilevanti disponibili come Linguistic Linked Open Data (LOD); l'altra applicazione è dedicata alla consultazione e interrogazione incrociata di questi dati. Questa integrazione mira a democratizzare l'accesso e la comprensione dei dati epigrafici e lessicografici, rendendoli fruibili da un vasto spettro di utenti, inclusi accademici, studenti e appassionati senza la necessità di competenze tecniche pregresse.

2.1. Architettura funzionale

La piattaforma DigItAnt adotta uno stile architetturale REST, caratterizzato da un'implementazione indipendente dei client e dei server. Nella Fig. 1 si vede un diagramma funzionale che descrive la struttura e le interazioni tra i vari componenti dello stack applicativo. All'interno di questo schema possiamo notare come la piattaforma di fruizione si colloca all'esterno dello stack, in quanto non deve passare attraverso altri servizi di autenticazione o altre componenti interne.
Le fonti di dati principali su cui si basa la piattaforma sono due: un servizio dedicato alle iscrizioni, rappresentato da CASH, che si occupa della gestione dei materiali epigrafici in TEI/XML, e un servizio per i lessici, rappresentato da LexO [1], un back-end che si occupa della gestione dei lessici modellati tramite lo standard OntoLex-Lemon. Questi software di back-end espongono API che operano sul protocollo HTTP e scambiano dati in formato JSON, conformandosi alle specifiche OpenAPI[6], un insieme di file d'interfaccia leggibili da macchine che descrivono, producono, consumano e visualizzano servizi REST.
Le funzionalità offerte dai servizi vengono attivate tramite due interfacce grafiche sviluppate in Angular, concepite come applicazioni web composte da diversi componenti. Ogni componente è dedicato a un insieme di diverse funzioni per la codifica o il recupero di vari aspetti delle voci lessicali e il loro collegamento con altre risorse dati sia interne che esterne, via URI, come voci di lessico, porzioni di testo, riferimenti bibliografici, vocabolari controllati e altre risorse LOD.

[4] Si tratta di lingue attestate tra VIII° secolo A.C e I° secolo D.C, risalente al periodo precedente alla romanizzazione della penisola italiana [5]. Per maggiori informazioni sulle lingue trattate si veda https://www.prin-italia-antica.unifi.it/p161.html
[5] https://github.com/EpiDoc/EFES
[6] https://www.openapis.org/

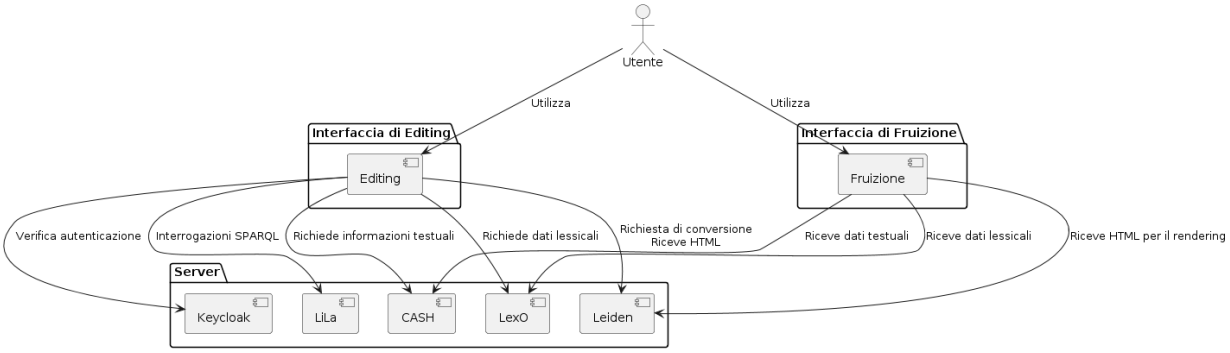


Figura 1. Schema funzionale dell'infrastruttura DigItAnt

2.2. Interfaccia di editing

L'interfaccia di editing consente di effettuare due tipologie di attività: la prima permette la creazione di un lessico conforme agli standard del Web Semantico basandosi su un corpus di testi epigrafici, mentre la seconda facilita il collegamento di elementi lessicali a token o porzioni di testo epigrafico. Nel primo caso, l'epigrafista importa un corpus di testi epigrafici in formato EpiDoc XML (già metadatato, ricostruito e opportunamente annotato) e inizia a compilare e codificare le voci lessicali attestate nel corpus, collegandole alle rispettive forme ricostruite nel testo, alla bibliografia pertinente e, ove possibile, a risorse esterne significative. Nel secondo caso, partendo da un lessico già esistente per le lingue di interesse, l'epigrafista importa un corpus e codifica i collegamenti ai vari dati, sia interni che esterni. Inoltre può modificare e arricchire il lessico aggiungendo nuovi elementi e nuove proprietà, come ad esempio le informazioni sull'etimologia di una parola grazie ad un'integrazione nel modello OntoLex-Lemon [4]. Come si può vedere nella figura 2, il testo dell'iscrizione appare in un componente presente nella sezione centrale dell'interfaccia, il quale è denominato come "Linker", che facilita l'associazione di porzioni di testo a voci del lessico attraverso l'uso dei servizi del server CASH.

Figura 2. Componente del Linker per collegare un elemento del testo ad una forma del lessico

Il processo di collegamento tra una porzione di testo e una forma lessicale corrisponde effettivamente alla creazione di una attestazione per la forma data.
Non entriamo qui nei dettagli di questa interfaccia e dei dati perché già descritti altrove, rispettivamente in [9, 10] e [6].

3. INTERFACCIA DI FRUIZIONE

Analogamente all'interfaccia di editing, quella di fruizione è concepita come un'applicazione web avanzata, progettata per visualizzare e interagire con i dati presenti negli endpoint di back-end, precedentemente creati o importanti. Questa applicazione si distingue per la sua capacità di integrare e presentare in modo coeso dati eterogenei, anche provenienti da fonti diverse, in particolar modo dati testuali rappresentati in XML con dati linguistici-lessicali in RDF (ospitati su database a grafo). L'interfaccia di front-end funge dunque anche da orchestratore delle interazioni tra le diverse sorgenti.

Una caratteristica distintiva di questa piattaforma di fruizione, illustrata in Fig. 3, è la sua strutturata divisione in aree funzionali, ciascuna dedicata a un aspetto specifico. La sezione dedicata ai testi e alle iscrizioni permette agli utenti di accedere a dettagliate trascrizioni testuali e a informazioni contestuali, come note storiche e geografiche, fornendo così una visione completa delle iscrizioni. Questa sezione si avvale di un componente che consente di visualizzare in maniera integrata il testo epigrafico, renderizzato secondo le convenzioni di Leida, insieme alle voci lessicali codificate nel dizionario computazionale collegato. Questo componente rappresenta un'innovazione rispetto a progetti di epigrafia digitale attualmente disponibili.

Parallelamente, la sezione dei lessici si rivela uno strumento importante per gli studi in prospettiva linguistica, permettendo di navigare tra voci e forme lessicali e di accedere immediatamente a informazioni approfondite su ogni elemento selezionato. Le informazioni lessicali sono anch'esse arricchite dalle informazioni testuali e bibliografiche codificate nelle risorse collegate e, quando disponibili e codificati, da collegamenti a risorse lessicali LOD esterne, come [7]. Questa focalizzazione sugli aspetti lessicali rappresenta un'altra innovazione nel panorama delle piattaforme oggi disponibili per l'epigrafia digitale e arricchisce la comprensione delle lingue arcaiche.

La bibliografia agisce da complemento essenziale, aggregando materiali bibliografici correlati e consentendo un ulteriore approfondimento degli studi.

Infine, la funzione di ricerca avanzata, che consente interrogazioni complesse e incrociate tra i diversi dataset, rappresenta un punto di forza della piattaforma.

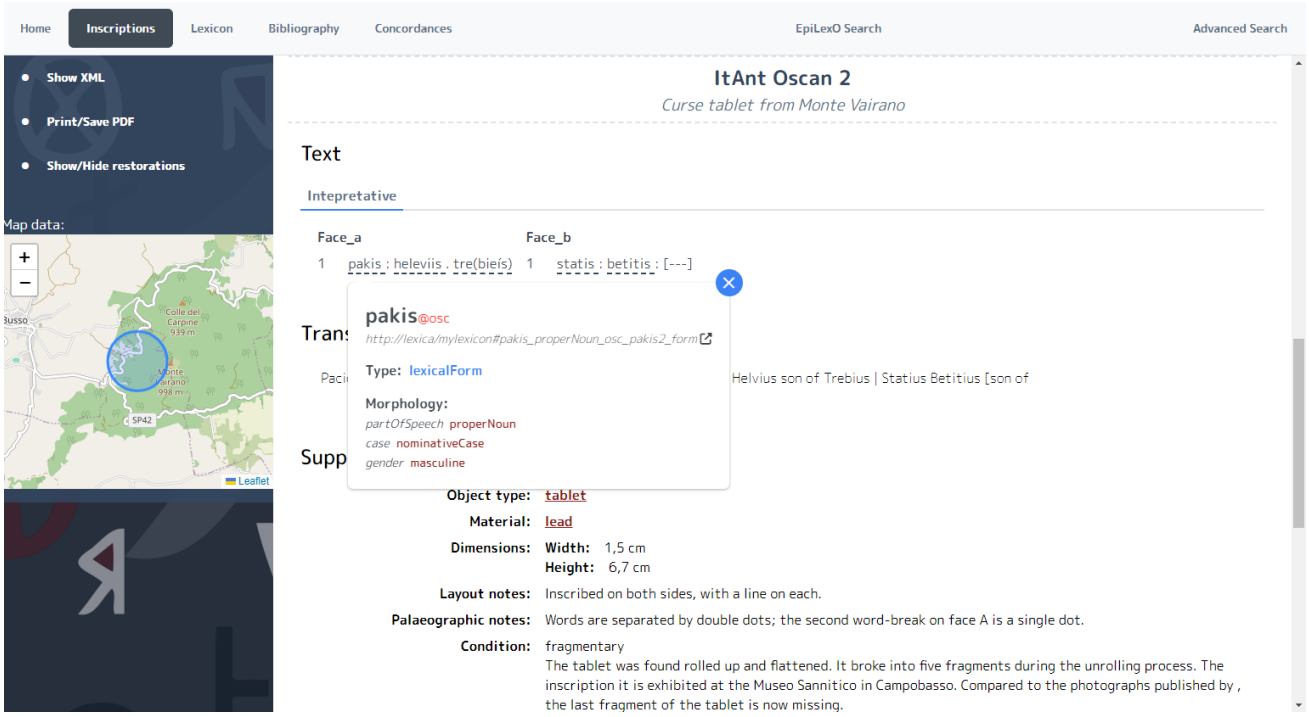


Figura 3. Vista dell'interfaccia di fruizione, sezione testi

La combinazione armoniosa di informazioni TEI/XML e LOD all'interno di questa interfaccia non solo migliora la qualità e la profondità delle informazioni presentate, ma rappresenta anche un salto qualitativo nell'esperienza utente, che può così avvalersi di una visione integrata e multiforme delle iscrizioni e dei lessici. L'integrazione dei LOD, in particolare, mette in risalto la capacità della piattaforma di collegare le informazioni lessicali e epigrafiche a una rete più ampia di conoscenze pre-esistenti, arricchendo l'esperienza di ricerca e di studio.

4. SOSTENIBILITÀ E INTEGRAZIONE INFRASTRUTTURALE

Poiché la piattaforma è stata sviluppata nel contesto di un progetto competitivo triennale, fin dalla sua progettazione è stato previsto un collegamento con l'infrastruttura CLARIN-IT, almeno per il deposito dei risultati digitali del progetto (dataset e software). Questo non solo a tutela dell'integrità del lavoro svolto, ma anche per facilitarne il riuso. CLARIN-IT offre infatti un servizio di deposito in un archivio digitale certificato[7] che garantisce, oltre alla conservazione e accessibilità a lungo termine, il versionamento e la citabilità delle risorse grazie all'assegnazione di identificativi univoci persistenti e alla catalogazione secondo metadati standardizzati altamente condivisi nella comunità scientifica.

Durante il progetto, si è deciso poi di trasferire anche la gestione della piattaforma DigItAnt a CLARIN-IT che potrà offrirla come servizio alla comunità, in modo da assicurare accessibilità e riutilizzo non solo dei dati, ma anche (del software) della piattaforma, anche per scopi diversi da quelli originari.

In questo senso DigItAnt rappresenta un caso di studio interessante, dal punto di vista della gestione tecnica dell'infrastruttura, per valutare la sostenibilità di questo tipo di servizi.

In questo paragrafo presentiamo quindi un'analisi preliminare, descrivendo le scelte infrastrutturali adottate e come queste possono essere migliorate in ottica di sostenibilità. La piattaforma, attualmente un prototipo operativo, risponde bene alle esigenze del progetto grazie a soluzioni software e hardware adeguate, con un accesso limitato ai collaboratori per le funzionalità di editing.

Per facilitare una transizione efficace verso un'infrastruttura più robusta, si è scelta l'adozione di Docker, che permetterà una migrazione fluida e scalabile attraverso la "containerizzazione" del software, garantendo la portabilità e l'indipendenza dal sistema operativo sottostante. Questo approccio facilita la gestione e il deployment delle applicazioni preparando la piattaforma per un aumento dell'utenza e delle richieste di accesso ai database[8]. Prevediamo che queste modifiche permetteranno alla piattaforma di sostenere un maggior numero di utenti e di gestire un volume molto più elevato di interrogazioni, consolidando la sua posizione come risorsa importante per lo studio delle lingue arcaiche dell'Italia pre-romana.

Per quanto riguarda l'utilizzo di dati, le attuali politiche di backup dell'applicazione di editing generano 12 MB di dati al giorno, con un lessico RDF che contiene 72.000 triple. Tali politiche non sono state (ancora) ottimizzate, ma potrebbero in caso aumento significativo di magnitudo. In ogni caso, la piattaforma, che occupa al momento 1 GB di spazio disco e richiede circa 4 GB di RAM per operare efficacemente, mostra una sostenibilità intrinseca dato che l'impatto delle risorse utilizzate non è eccessivo per l'infrastruttura ospitante. Anche se il progetto è quasi concluso, si prevede che la creazione e l'aggiunta di nuovi dati continuerà anche oltre il termine anche se con un ritmo diverso. In ogni caso, data la natura epigrafica frammentaria dei materiali primari e della forte specialità dell'ambito disciplinare, la quantità rimarrà naturalmente contenuta. L'impegno richiesto per mantenere la piattaforma web di fruizione è quindi gestibile, rendendo la manutenzione dell'infrastruttura praticabile ed economicamente sostenibile. Con i parametri di spazio attuali e con le quantità di dati attualmente disponibili[9], pensiamo che si possano ospitare molte altre applicazioni che non contengano dati pesanti, come ad esempio immagini ad alta risoluzione o eventualmente altri dati multimediali che occupano un certo spazio; in quest'ultimo caso, andrebbero adottate delle strategie diverse. In questo scenario, l'elevata efficienza operativa si traduce in vantaggi sia tecnologici che economici, assicurando una gestione dei dati duratura e affidabile. Sebbene la strategia di versionamento e conservazione assicuri la disponibilità e l'integrità dei dati nel tempo, è importante sottolineare che la correlazione diretta con l'accessibilità dei dati attraverso la piattaforma può non essere immediata. I dati, conservati e versionati nel repository, richiedono un'interfaccia separata per l'accesso e l'utilizzo effettivo. Pertanto, mentre il versionamento garantisce la preservazione dei dati, l'accessibilità e l'utilizzo pratico necessitano di soluzioni aggiuntive per garantire la piena efficacia della piattaforma, riutilizzabili e conformi ai principi FAIR, favorendo così una maggiore sostenibilità e utilità a lungo termine del progetto.

5. CONCLUSIONI

L'articolo ha esplorato l'implementazione e l'integrazione di una piattaforma di creazione e fruizione sviluppata nel contesto di un progetto di ricerca triennale, sottolineando l'importanza di collegamenti strategici con infrastrutture esistenti come CLARIN-IT. Questo non solo garantisce la conservazione e la citabilità a lungo termine dei dati raccolti, ma supporta anche la loro disseminazione e il riuso all'interno della comunità scientifica.

[7] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/

[8] Finora, la piattaforma non ha affrontato un uso massivo. Un aumento significativo dell'uso potrebbe richiedere adattamenti e ottimizzazioni. Tuttavia, l'accesso a quest'ultima limitato ai soli studiosi autorizzati, e la natura di nicchia del progetto riducono il rischio di sovraccarico. Le robuste tecnologie di back-end mitigano ulteriori rischi per l'interfaccia di fruizione, rendendo improbabili problemi legati a un uso massimo simultaneo.

[9] Si prevede che entro la fine del progetto i dati aumenteranno, ma l'ordine di grandezza non si sposterà di molto.

Per quanto riguarda le prospettive future, la piattaforma DigItAnt continuerà a operare sotto l'egida di CLARIN-IT come una "vetrina" stabile e affidabile per i dati digitali delle lingue dell'Italia Antica. Rimarrà attiva anche nella modalità di editing per gli utenti autorizzati, permettendo l'aggiornamento e l'espansione dei dataset, in particolare dei corpora e dei lessici relativi anche ad altre lingue epigrafiche frammentarie.
Al momento non sono previste innovazioni o cambiamenti sostanziali alla piattaforma, il cui valore intrinseco risiede nella capacità di presentare un'interfaccia coerente e intuitiva per la gestione e interrogazione di dati linguistici collegati. In futuro, gli sviluppi potrebbero includere l'ottimizzazione delle funzionalità esistenti e l'ampliamento delle capacità di integrazione dei Linked Open Data, per migliorare l'interoperabilità e arricchire ulteriormente le connessioni tra diversi set di dati. Questo rafforzamento potrebbe facilitare una comprensione più profonda e una maggiore fruibilità delle informazioni in ambito epigrafico e linguistico delle lingue antiche.

6. RINGRAZIAMENTI

Questo lavoro è stato svolto nel contesto dei seguenti progetti: H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 4 "Istruzione e Ricerca" Componente 2 "Dalla ricerca all'impresa" Linea di Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione" Azione 3.1.1 "Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti" - Codice progetto IR0000029 - CUP B63C22000730005. Soggetto attuatore CNR; e PRIN 2017XJLE8J "Lingue e Culture dell'Italia antica. Linguistica Storica e Modelli Digitali".

BIBLIOGRAFIA

[1] Bellandi, Andrea. «LexO: an open-source system for managing OntoLex-Lemon resources». *Language Resources and Evaluation* 55 (2021): 1093–1126. https://doi.org/10.1007/s10579-021-09546-4.

[2] Da Sylva, Lyne. «Towards Linked Data: Some Consequences for Researchers in the Social Sciences and Humanities». In *Proceedings of the Association for Information Science and Technology*, a cura di Luanne Freund, 94–103. Hoboken, NJ: Wiley, 2018.

[3] Hawkins, Ashleigh. «Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-Digital Archives via the Semantic Web». *Archival Science* 22, fasc. 3 (settembre 2022): 319–44. https://doi.org/10.1007/s10502-021-09381-0.

[4] Khan, Anas F. «Towards the Representation of Etymological Data on the Semantic Web». *Information* 9, fasc. 12: 304 (2018): 1–17. https://doi:10.3390/info9120304.

[5] Marinetti, Anna. «Scritture e lingue dell'Italia antica». In *Le grandi vie della civiltà. Relazioni e scambi fra Mediterraneo e il Centro Europa dalla preistoria alla romanità*, a cura di Franco Marzatico, 385–91. Trento: Castello del Buonconsiglio. Monumenti e collezioni provinciali, 2011.

[6] Murano, Francesca, Valeria Quochi, Angelo Mario Del Grosso, Luca Rigobianco, e Mariarosaria Zinzi. «Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process». *Journal on Computing and Cultural Heritage* 16.3, fasc. 1 (2023).

[7] Passarotti, Marco C., e Francesco Mambrini. «Linking Latin: Interoperable Lexical Resources in the LiLa Project». In *Building new resources for historical linguistics*, a cura di Erica Biagetti, Chiara Zanchi, e Silvia Luraghi, 103–24. Pavia: Pavia University Press, 2021. https://hdl.handle.net/10807/194955.

[8] Prag, Jonathan R.W., James Chartrand, e James Cummings. «I. Sicily: an EpiDoc corpus for ancient Sicily». In *Digital and Traditional Epigraphy in Context: The Proceedings of the Second EAGLE International Conference*, a cura di Silvia Orlandi, Raffaella Santucci, Francesco Mambrini, e Pietro Maria Liuzzo. Newcastle: Newcastle University, 2016.

[9] Quochi, Valeria, Andrea Bellandi, Anas F. Khan, Michele Mallia, Francesca Murano, Silvia Piccini, Luca Rigobianco, Alessandro Tommasi, e Cesare Zavattari. «From Inscriptions to Lexica and Back: A Platform for Editing and Linking the Languages of Ancient Italy». In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, a cura di Rachele Sprugnoli e Marco Passarotti, 59–67. Marseille: European Language Resources Association, 2022. https://aclanthology.org/2022.lt4hala-1.9.

[10] Quochi, Valeria, Andrea Bellandi, Michele Mallia, Alessandro Tommasi, e Cesare Zavattari. «Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO». In *CLARIN Annual Conference Proceedings*, a cura di Tomaž Erjavec e Maria Eskevich, 39-43. Prague: Czechia, 2022.

[11] Rigobianco, Luca. «La linguistica delle lingue di attestazione frammentaria». In *Metodi e prospettive della ricerca linguistica*, a cura di Chiara Meluzzi e Nicholas Nese, 29:83–94. Ledizioni, 2022. https://iris.unive.it/handle/10278/3762809.

[12] Vagionakis, Irene. «Cretan Institutional Inscriptions: A New EpiDoc Database». *Journal of the Text Encoding Initiative*, 2021, 1–21. https://journals.openedition.org/jtei/3570.

Infrastrutture di ricerca come strumenti di "interculturalità digitale"

Salvatore Cristofaro[1], Vittoria Fabiani[2], Cristina Marras[3],
Enrico Pasini[4], Pietro Sichera[5], Mingyang Yu[6]
[1]CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italia - salvatore.cristofaro@cnr.it
[2]CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italia - vittoria.fabiani@cnr.it
[3]CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italia - cristina.marras@cnr.it
[4]CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee / Università di Torino, Italia - enrico.pasini@cnr.it
[5]CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee, Italia - pietro.sichera@cnr.it
[6]Università di Torino, Italia - mingyang.yu@edu.unito.it

ABSTRACT

Il contributo si focalizza su alcune caratteristiche delle infrastrutture di ricerca (IR) per le scienze umane e il patrimonio culturale (SSH) che facilitano la scienza aperta. In particolare, presenta alcune funzioni di un marketplace per le SSH come esempio di messa in relazione e integrazione di strumenti per aggregare, accedere, rappresentare, condividere, riusare e comunicare la conoscenza. L'articolo è organizzato in tre parti: nella prima parte, si introducono le infrastrutture di ricerca come strumenti interculturali e interdisciplinari soffermandosi nello specifico su OPERAS (OPen Scholarly Communication in the European Research Area for Social Sciences and Humanities) e sul nodo italiano di recente costituzione. Nella seconda parte si presentano alcuni dei requisiti di design (*back-end* e *front-end*), i workflow di lavoro e le politiche di accesso per un marketplace che costituisca un punto di aggregazione per infrastrutture federate quale quello in fase di sviluppo nel progetto H2IOSC (Humanities and Heritage Italian Open Science Cloud). La terza parte riflette sul marketplace inteso non solo come vetrina di offerta di servizi ma come uno spazio 'transdisciplinare' e 'transnazionale', in cui sia nella fase di progettazione sia nell'utilizzo si possano sperimentare e mettere in pratica le potenzialità collaborative della ricerca digitale, in cui è all'opera una forma di "interculturalità digitale".

PAROLE CHIAVE

Infrastrutture, Open Science, Riusabilità, Sostenibilità, Scholarly Communication

1. INFRASTRUTTURE DIGITALI: ECOSISTEMI DI RICERCA PER LA SCIENZA APERTA

Le infrastrutture digitali[1] creano un ecosistema di ricerca e si caratterizzano per iniziative transdisciplinari; mettono insieme infatti specialisti di dati, analisti, ingegneri, ricercatori, creando ponti da e verso altre infrastrutture esistenti e immaginando ambienti di ricerca open source modulari, promuovendo così la trasformazione digitale nella ricerca e il trasferimento di competenze e conoscenze [3, 5].

La comunicazione scientifica nelle SSH ha sofferto di frammentarietà e di una certa difficoltà nel passaggio verso l'Open Science (OS). Questi problemi si sono evidenziati a causa di diversi fattori, primo fra tutti la varietà culturale della comunità scientifica, senza dimenticare o sottovalutare la spesso ridotta dimensione delle risorse (nella doppia valenza semantica di risorse condivise/risorse economiche).

Nel panorama delle infrastrutture digitali, OPERAS[2] fornisce una piattaforma di servizi innovativi dedicati alla comunicazione scientifica aperta nelle SSH. Si tratta di una infrastruttura di ricerca distribuita, operante nello spazio europeo della ricerca[3], che affronta alcune sfide specifiche, come la diversità linguistica nelle pubblicazioni e nell'accesso ai contenuti, il libero accesso e la gestione di forme specifiche della produzione scientifica (monografie, edizioni critiche, ecc.), l'accesso aperto "Diamond". OPERAS intende realizzare un sistema di comunicazione scientifica aperto, senza barriere, che consenta alla comunità di ricerca SSH di coordinare e federare le risorse, e di reperire, consultare, creare, modificare, diffondere e convalidare i risultati della ricerca in tutta Europa in modo semplice ed efficiente. Operativamente,

[1] Cfr. https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/european-research-infrastructures_en
[2] https://roadmap2021.esfri.eu/projects-and-landmarks/browse-the-catalogue/operas/
[3] OPERAS, selezionata come infrastruttura di ricerca con ruolo chiave in Europa dalla Roadmap 2021 dell'ESFRI, ha sede a Bruxelles e conta attualmente 63 membri, per lo più provenienti da paesi europei, tra cui il Regno Unito. Attualmente ha la forma legale di un'associazione internazionale senza fini di lucro e ha avviato il processo di trasformazione in ERIC (consorzio europeo di infrastruttura di ricerca) che si prevede giunga a compimento nel 2028 (https://operas-eu.org).

si tratta di assicurare in diverse forme e con diversi strumenti (come per esempio Go-Triple[4]), l'accesso transnazionale alle risorse e ai servizi di comunicazione accademica disponibili per chi fa ricerca e integrare gli strumenti nell'Open Science Cloud europea (EOSC)[5]. Intorno a OPERAS si muovono infatti numerosi progetti e iniziative (OPERAS-PLUS, COESO, TRIPLE, DIAMAS, CO-OPERAS)[6] che hanno determinato la costituzione di vari gruppi di lavoro e nuove comunità scientifiche e di collaborazione (tanto che si parla di *OPERAS Universe*). OPERAS ridisegna il modo di considerare la ricerca nelle SSH perché, pur seguendo un processo di standardizzazione e normalizzazione (per esempio nella metadatazione delle risorse) non pensa di appiattire, bensì di esaltare le peculiarità scientifiche di ogni contenuto, garantendone, in linea con i principi FAIR[7] [7] e TRUST[8] [6], l'accessibilità, la rintracciabilità, l'interoperabilità e la sostenibilità.

Di recente si è costituito il nodo italiano OPERAS-IT di OPERAS, una Joint Research Unit (OPERAS.it JRU), coordinata dal Consiglio Nazionale delle Ricerche. Già dal 2020 OPERAS è stata inserita nel Piano nazionale delle infrastrutture di ricerca (PNIR) con alta priorità, e insieme con i nodi italiani di CLARIN[9], DARIAH[10], E-RIHS[11] partecipa al progetto H2IOSC[12]. L'obiettivo del progetto è quello di creare una Open Science Cloud italiana per la ricerca nelle scienze umanistiche, linguistiche e del patrimonio culturale, incentivando un approccio aperto e FAIR a dati, strumenti, documentazione della ricerca stessa.

Il progetto H2IOSC[13] è suddiviso in 8 work-package (WP), ciascuno con diversi ambiti di azione, ma operanti in sinergia gli uni con gli altri (vd. Fig. 1).

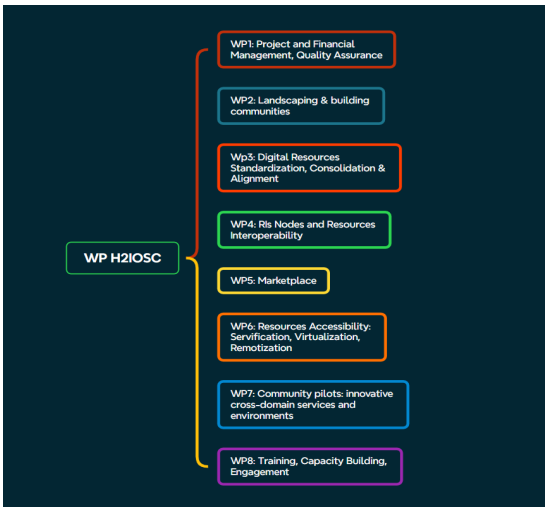


Figura 1. Work-package di H2IOSC

Nell'ambito delle diverse attività di progetto, OPERAS è responsabile tra l'altro del coordinamento del WP5, nel quale viene sviluppato il Marketplace, la 'vetrina' di accesso ai servizi sviluppati da H2IOSC, in cui si potranno altresì rendere disponibili cataloghi e risorse provenienti sia dalle IR italiane sia da altri nodi nazionali e controparti europee[14]. I WP direttamente connessi con il Marketplace (vd. Fig. 2) ambiscono, in particolare, a:

- assicurare il funzionamento della piattaforma che sarà sviluppata dal progetto e la sua coerenza formale e progettuale;
- fornire servizi scientifici e repository specifici per i servizi sviluppati specificamente (piloti) o virtualizzati dalle IR partecipanti;

[4] https://www.gotriple.eu/
[5] https://eosc-portal.eu/
[6] https://operas-eu.org/projects/
[7] FAIR Findable, Accessible, Interoperable, and Reusable, cfr. https://www.go-fair.org/fair-principles/
[8] TRUST, Transparency, Responsibility, User focus, Sustainability and Technology.
[9] https://www.clarin.eu/
[10] https://www.dariah.eu/
[11] https://www.e-rihs.eu/
[12] https://www.h2iosc.cnr.it/
[13] https://www.h2iosc.cnr.it/home/
[14] OPERAS-IT coordina anche il WP7 (Community pilots: innovative cross-domain services and environments) che ha l'obiettivo di definire e realizzare delle *applicazioni pilota*, servizi orientati all'innovazione e a risorse sperimentali, implementate come "proof of concept".

- predisporre un servizio di training per gli utenti della piattaforma;
- attivare scambi di mobilità transnazionale con le IR europee di riferimento.

Il Marketplace di H2OISC rappresenta il luogo di facile accesso dove gli utenti potranno fruire di numerosi servizi, strumenti, dataset, software, separatamente o aggregati in workflow, a sostegno delle loro specifiche esigenze di ricerca.

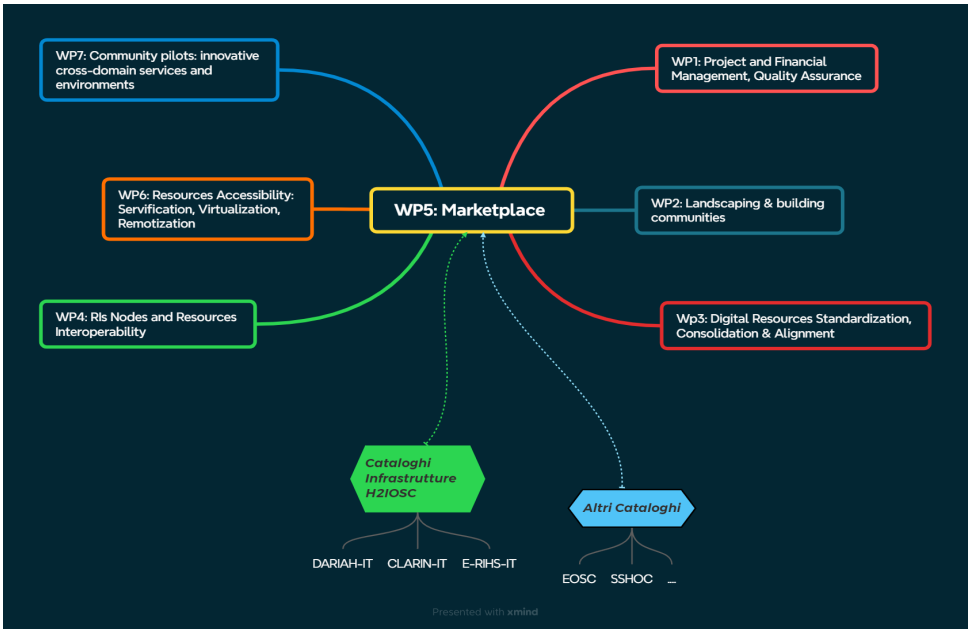


Figura 2. Il Marketplace di H2IOSC come vetrina di servizi e aggregatore di attività

2. MARKETPLACE

La formula dei 'marketplace' è stata identificata nello spazio della ricerca europeo come uno strumento essenziale per sostenere la transizione dalla ricerca ordinaria alle infrastrutture basate sul cloud. In termini molto generali, un marketplace rappresenta un 'luogo' dove vengono scambiati beni e servizi. Nel particolare contesto delle SSH, tali beni e servizi possono includere (e coinvolgere), allo stato attuale, grandi e complessi moli di dati (strutturati ed eterogenei) relativi a ricerche in ambito umanistico, sociale e culturale, e strumenti software per l'interrogazione, l'analisi e quindi la fruizione di questi stessi dati. L'evoluzione delle tecnologie di comunicazione ha prodotto un massiccio aumento del numero delle persone che pubblicano e scambiano sul WEB, anche grazie allo sviluppo dei moderni Cloud Networks che favoriscono la gestione dei dati in maniera condivisa e collaborativa (Open Data) offrendo potenti piattaforme di calcolo computazionale e di immagazzinamento delle informazioni. In questo contesto, infatti, il paradigma dell'Open Cloud Computing si è affermato come strumento predominante per la distribuzione di infrastrutture e risorse di rete, bypassando la necessità di disporre in loco di complessi e costosi sistemi hardware/software per la gestione (anche collaborativa) dei dati. Un marketplace può fungere da punto di raccordo tra diversi Cloud Networks e altre infrastrutture di rete, realizzando un ambiente di accesso unico, dinamico e funzionale, per la condivisione, la fruizione e la gestione collaborativa di dati e servizi.

In quest'ottica anche il Marketplace di H2IOSC ha come obiettivo quello di aumentare la visibilità e la valorizzazione delle fonti di dati, dei servizi e delle risorse fornite dalle IR confederate nel progetto e dalla comunità di ricerca in generale, ospitate nel Cloud Nazionale H2IOSC in piena conformità con la normativa vigente e nel pieno rispetto dei principi FAIR. Pubblicizza strumenti e metodologie, introduce strumenti di collaborazione e permette ai partecipanti di condividere e rendere reperibili dataset e strumenti di ricerca, nonché materiali di formazione, cataloghi contestualizzati, strumenti di presentazione, visualizzazione e aggregazione. Crea e favorisce collegamenti tra persone, dati, servizi e formazione.

L'architettura del Marketplace comprende una piattaforma con una struttura informatica per favorire la visibilità e l'accessibilità dei contributi dei fornitori dei contenuti, e un insieme di funzionalità per l'interazione da parte degli utenti (vd. Fig. 3). Ciò implementa il concetto di 'mercato', nell'accezione più generale possibile del termine, come luogo dove facilmente è possibile far incontrare domanda e offerta. Il mercato è spazio fisico (in questo caso digitale), gestito, riservato, orientato alla fruizione migliore possibile del servizio, luogo di acquisizione diretta di offerte molteplici. Questo modello, legato alla quotidianità, lo ritroviamo nelle grandi vetrine online: chi vende può offrire articoli singoli, così come può organizzare dei veri e propri negozi, in cui mostrare i propri prodotti in maniera dettagliata, tematica, organica, esplicitando diverse caratteristiche e particolarità. Può inoltre utilizzare le funzioni automatiche (API - Application Program Interface), che consentono di effettuare delle operazioni sul proprio negozio in maniera non presidiata, come ad esempio conoscere

l'elenco dei prodotti attualmente offerti o aggiornare le quantità messe in vendita. L'utente può effettuare ricerche secondo vari parametri, potendo valutare l'affidabilità dei risultati e raffinarli progressivamente. Il Marketplace di H2IOSC, come già avviene nei suoi modelli a livello europeo, porta nel mondo della ricerca aperta e collaborativa questi elementi metaforici e di interazione. Da un punto di vista tecnico, è progettato per offrire delle funzionalità su diversi livelli, in modo da soddisfare le esigenze tanto degli amministratori quanto degli utenti. Il livello più basso (*livello 0*) offre i servizi costitutivi del Marketplace (non inclusi nel catalogo), che comprendono:

- funzionalità orientate all'utente (navigazione, liste personali e condivisibili dei "preferiti"...);
- funzionalità orientate alla governance (amministrazione, statistiche...).
- Al livello successivo (*livello 1*) si trovano i servizi offerti dal catalogo del Marketplace. Consideriamo come servizi di livello 1:
- l'interfaccia di ricerca che guida gli utenti nel reperire le voci di loro interesse;
- i servizi esposti dal Marketplace.

Il catalogo non conterrà soltanto i servizi sviluppati o virtualizzati espressamente per H2IOSC: sarà popolato anche tramite importazione sia automatica sia manuale di dati esterni e, in particolare, garantirà l'interoperabilità con i marketplace di EOSC[15] e SSHOC[16]. Ugualmente, cataloghi esterni potranno accedere al Marketplace di H2IOSC, tramite protocolli standard per l'harvesting e l'esposizione di API dedicate.
Il livello superiore (*livello 2*) è il livello di aggregazione dei servizi complessi (*servizi pilota*), come per esempio le diverse funzionalità per un workflow di lavoro per le *Scholarly Digital Editions* (codifica, layout, fairificazione…), alcuni dei quali sono implementati da parte degli altri WP del progetto H2IOSC. Vi saranno dunque API che definiscono le regole e i formati con cui un'applicazione (risorsa o servizio di livello 1) può interagire con un'altra, consentendo di integrarle e scambiare informazioni in modo efficace senza doverne conoscere i dettagli interni. Particolare attenzione è dedicata alla modellazione dei dati e, logicamente, al supporto ai principi FAIR che il Marketplace (e il cloud H2IOSC in generale) deve garantire.

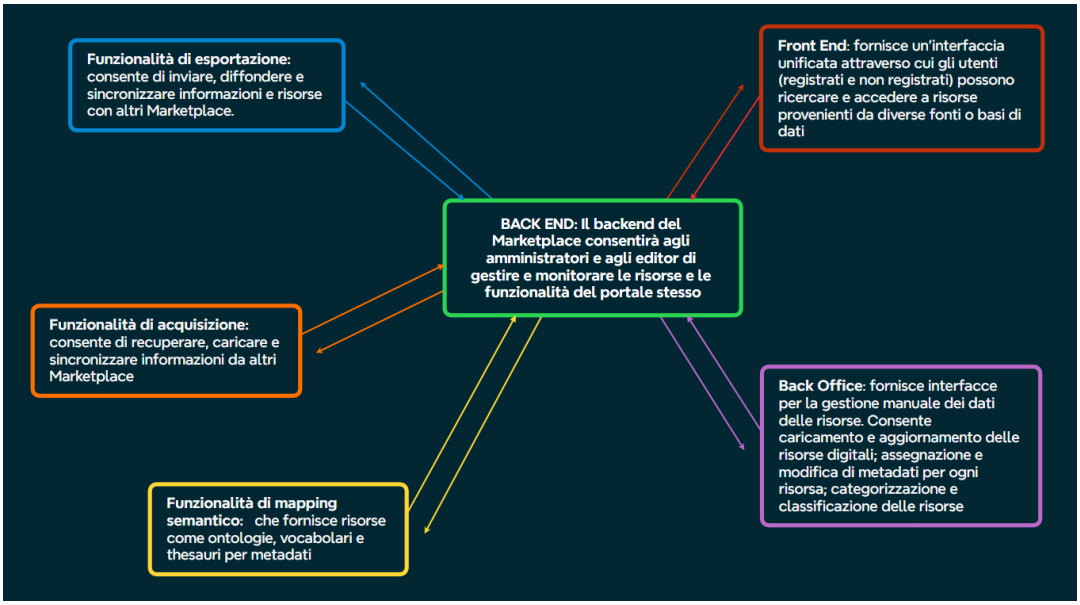


Figura 3. Architettura tecnologica per lo sviluppo del Marketplace.

3. INTERCULTURALITÀ DIGITALE

L'interculturalità è l'instaurazione e il mantenimento di rapporti culturali come forme di dialogo, di confronto e di reciproco scambio di conoscenze e di prospettive, e rappresenta un affascinante e complesso intreccio di diversità che va oltre la mera coesistenza [1]. È un processo dinamico in cui le varie espressioni culturali si fondono in un dialogo fruttuoso, arricchendosi reciprocamente attraverso confronti profondi [4]; in ambito digitale è un modello teorico che ha bisogno di strumenti concreti per essere attuato in maniera consistente e significativa.
Da questo punto di vista, è evidente che le infrastrutture dello spazio europeo della ricerca come OPERAS, CLARIN, DARIAH, E-RIHS fungono da ponti tra mondi e tradizioni scientifiche culturali diverse. Ciascuna di esse, infatti, è

[15] Per il catalogo di EOSC si veda: https://marketplace.eosc-portal.eu/
[16] Per il catalogo di SSHOC si veda: https://www.sshopencloud.eu/ssh-open-marketplace

partecipata da diversi paesi europei e opera in collaborazione con altri stati membri e partner istituzionali nazionali. Esse sono, allo stesso tempo, anche luoghi che traducono concretamente l'idea di interculturalità in azione tangibile. I loro servizi per la comunità scientifica affrontano, per esempio, i temi dell'educazione e della formazione digitale, la frammentarietà delle risorse e la loro spesso difficile accessibilità (si pensi alle azioni di *open humanitie*s e *public humanitie*s o *Arts exchange* di DARIAH). Hanno al proprio centro la questione delle lingue e del multilinguismo, come si vede molto bene nei servizi di CLARIN. Si dedicano specificamente alla memoria storica e al patrimonio culturale (E-RIHS). I principi FAIR e TRUST a cui si rifanno e che possono rappresentare uno dei principi fondamentali della loro attività (OPERAS) costituiscono linee guida di riferimento, nel rispetto delle specificità disciplinari, per garantire uno spazio di 'fiducia scientifica'[17] libero, aperto e affidabile per la ricerca.

H2IOSC, che riunendo i nodi italiani delle infrastrutture europee ne assume i principi guida e gli ambiti operativi, è esso stesso uno spazio interculturale aperto e condiviso di risorse, servizi e conoscenza. In questo laboratorio sperimentale, il Marketplace costituirà la vera e propria piazza del mercato (l'*agorà* che dal tempo della città greca rappresenta un luogo sociale primario di scambio e incontro nella costituzione dello spazio politico democratico) dove avverranno i principali processi di scambio e interscambio scientifico e culturale [2]. È, per questo, ben più di una vetrina o un catalogo: è uno strumento che veicola le forme di aggregazione transdisciplinare e lo sviluppo di conoscenza e di tecnologie, facilitatore dunque della creazione di strumenti, metodi e processi comuni e condivisi: virtuali sì, ma autenticamente plurali, in cui operano nuovi modelli di "interdisciplinarità digitale".

BIBLIOGRAFIA

[1] Cassano, Franco. *Il pensiero meridiano*. Roma-Bari: Laterza, 2007.

[2] Conrad, Arensberg M., Harry W. Pearson, e Karl Polanyi. *Traffici e mercati negli antichi imperi: le economie nella storia e nella teoria*. Torino: Einaudi, 1978.

[3] Del Rio Riande, Gimena. «Digital Humanities and Visible and Invisible Infrastructures». In *Global debates in the Digital Humanities*, a cura di Domenico Fiormonte, Sukanta Chaudhari, e Paola Ricaurte. University of Minnesota Press, 2022. https://dhdebates.gc.cuny.edu/projects/global-debates-in-the-digital-humanities.

[4] Hall, Edward T., e Mildred Reed Hall. *Understanding cultural differences*. Intercultural Press, 1990.

[5] Liburdi, Annarita, Cristina Marras, e Ada Russo. «Infrastrutture, terminologie e policy per la ricerca umanistica: note per un confronto interdisciplinare». In *Conferenza GARR 2018 - Data (R)evolution - Selected Papers*. Cagliari, 2018. https://www.garr.it/it/chi-siamo/documenti/selected-papers/selected-papers-conferenza-2018/4711-selected-papers-conferenza-2018-15-liburdi/file.

[6] Lin, Dawei, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, et al. «The TRUST Principles for digital repositories». *Scientific Data* 7 (2020): Article number: 144. https://doi.org/10.1038/s41597-020-0486-7.

[7] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3, fasc. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[17] Cfr. https://cordis.europa.eu/project/id/872855

Materiali didattici come oggetti digitali FAIR: una metodologia condivisa per la formazione in H2IOSC

Giulia Pedonese[1], Francesca Frontini[2], Roberta Ottaviani [3], Federico Boschetti [4], Alessia Spadi[5], Lucia Francalanci[6], Alessia Scognamiglio[7], Pietro Restaneo[8], Antonina Chaban[9], Jana Striova[10], Laura Benassi[11]

[1]CNR Istituto di Linguistica Computazionale "Antonio Zampolli", Italia – giulia.pedonese@cnr.it
[2]CNR Istituto di Linguistica Computazionale "Antonio Zampolli", Italia – francesca.frontini@cnr.it
[3]CNR Istituto di Linguistica Computazionale "Antonio Zampolli", Italia – roberta.ottaviani@cnr.it
[4]CNR Istituto di Linguistica Computazionale "Antonio Zampolli", Italia – federico.boschetti@cnr.it
[5]CNR Istituto Opera del Vocabolario Italiano, Italia – alessia.spadi@cnr.it
[6]CNR Istituto Opera del Vocabolario Italiano, Italia – lucia.francalanci@cnr.it
[7]CNR Istituto per la Storia del Pensiero Filosofico e Scientifico Moderno, Italia – alessia.scognamiglio@cnr.it
[8]CNR Istituto per il Lessico Intellettuale Europeo e Storia, delle Idee, Italia – pietro.restaneo@cnr.it
[9]CNR Istituto Nazionale di Ottica, Italia – antonina.chaban@cnr.it
[10]CNR Istituto Nazionale di Ottica, Italia – jana.striova@cnr.it
[11]CNR Istituto Nazionale di Ottica, Italia – laura.benassi@cnr.it

ABSTRACT

Il presente lavoro dettaglia la strategia per lo sviluppo di iniziative di formazione nell'ambito del progetto H2IOSC e mira a coinvolgere la comunità italiana di riferimento sulle modalità di design e di fruizione di moduli didattici che integrino l'uso delle Infrastrutture di Ricerca. In particolare, il contributo si sofferma sulla descrizione dei requisiti per l'implementazione dell'infrastruttura di *training* e sugli standard condivisi per la descrizione dei materiali didattici come oggetti digitali FAIR al fine di massimizzarne il riutilizzo in un'ottica *train the trainers.*

PAROLE CHIAVE

Formazione; training; infrastrutture di ricerca; H2IOSC; principi FAIR.

1. FORMAZIONE NELLE INFRASTRUTTURE DI RICERCA

Il progetto Humanities and cultural Heritage Italian Open Science Cloud (H2IOSC) [1] intende creare un cluster federato dei servizi e delle risorse sviluppate dai nodi nazionali di quattro Infrastrutture di Ricerca (IR) per la Scienza Aperta[1] che fanno parte della *roadmap* European Strategy Forum on Research Infrastructure (ESFRI)[2] nel settore dell'innovazione sociale e culturale: Digital Research Infrastructures for the Arts and Humanities (DARIAH); European Research Infrastructure for Heritage Science (E-RIHS), Common Language Resource and Technology Infrastructure (CLARIN) e Open Scholarly Communication in the European Research Area for Social Sciences and Humanities (OPERAS)[3]. Il presente contributo ha l'obiettivo di aprire la strategia di formazione sviluppata nell'ambito del progetto H2IOSC alla comunità scientifica in modo da favorire l'interscambio di obiettivi, metodologie e strumenti didattici volti ad integrare l'uso delle IR all'interno dei programmi di istruzione superiore, universitaria e professionale.

Per il funzionamento delle IR intese come «strutture, risorse e servizi correlati che sono utilizzati dalla comunità scientifica per condurre ricerche di alto livello nei rispettivi campi del sapere» [3: 197] la formazione è un aspetto cruciale in quanto costituisce un ponte fra le comunità scientifiche e ciò che le IR offrono, risultando strettamente connessa alle attività di disseminazione che promuovono l'uso di prodotti e servizi da parte di ricercatori, studenti e cittadini. Più in generale, le IR contribuiscono all'istruzione fornendo una formazione specializzata a studenti, ricercatori e professionisti del mondo accademico e industriale su metodi e tecnologie scientifiche all'avanguardia. Per questo le iniziative di formazione hanno da sempre un posto di primo piano attraverso l'organizzazione di eventi come seminari, *workshop* e *summer school*, l'erogazione di borse di studio e di mobilità e la creazione di archivi digitali attraverso cui gli utenti possono accedere al

[1] https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en
[2] https://roadmap2021.esfri.eu/landscape-analysis/section-1/social-cultural-innovation/
[3] H2IOSC rientra nell'ambito del PNRR Missione 4, "Istruzione e Ricerca" - Componente 2, "Dalla ricerca all'impresa" - Linea di investimento 3.1, "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione", finanziato dall'Unione europea – programma NextGenerationEU, di cui al Decreto Direttoriale MUR n. 0003264 del 28/12/2021 codice progetto IR0000029, CUP B63C22000730005. Sito ufficiale: https://www.h2iosc.cnr.it/

materiale didattico, indicizzato secondo standard condivisi in quanto oggetto digitale in aderenza ai principi FAIR[4]. A questo proposito si ricordano non solo le iniziative di infrastrutture disciplinari come CLARIN (con il suo Learning Hub) e DARIAH (con la piattaforma DARIAH-Campus[5]), che forniscono l'accesso a risorse formative quali registrazioni di eventi, slide deck e tutorial indicizzati per tipologia di argomento o evento, ma anche le risorse dedicate al *training* da grandi progetti cluster europei fra i quali si segnala il Training Discovery Toolkit[6] sviluppato nell'ambito del Social Sciences and Humanities Open Cloud[7] (SSHOC). Inoltre, per potenziare le competenze trasversali degli utenti in termini di principi FAIR, Scienza Aperta e gestione dei dati della ricerca, il progetto Skills for the European Open Science commons: creating a training ecosystem for Open and FAIR science (Skills4EOSC) finanziato dal programma Horizon Europe della Commissione Europea e coordinato dal Consorzio GARR, ricopre un ruolo d'avanguardia nello sviluppo di metodologie, attività e risorse condivise per unificare la formazione dei professionisti della ricerca che costituiranno il futuro bacino d'utenza dello European Open Science Cloud[8] (EOSC). Il progetto H2IOSC si inserisce pienamente in questo contesto dedicando un intero *work package* (WP) alla formazione e basandosi sulle iniziative preesistenti per la Scienza Aperta e la gestione di dati FAIR per concentrarsi sull'ambito delle Social Sciences and Humanities (SSH) con casi d'uso specifici per le comunità afferenti.

2. LA STRATEGIA DI TRAINING H2IOSC

Il WP dedicato alla formazione è rivolto alle comunità di utenti italiani con lo scopo di dotarli di capacità e competenze interdisciplinari specifiche del settore SSH e formarli sulle risorse che le infrastrutture disciplinari possono offrire a livello nazionale e internazionale. Parte di questo obiettivo è quello di formare nuove figure professionali in grado di formare a loro volta le future generazioni su come integrare le IR nei metodi e nelle pratiche di ricerca delle rispettive discipline, con una prospettiva *train the trainers,* da cui l'esigenza di disporre di materiali didattici modulari e facilmente riutilizzabili da poter essere integrati nei propri corsi di formazione a seconda delle necessità. Le unità operative coinvolte nel WP hanno sviluppato una strategia condivisa per il coordinamento delle attività di *training*, sia a livello di singole IR che di progetto, che comprende innanzitutto l'individuazione dei bisogni formativi delle diverse tipologie di utenti, l'implementazione di un'infrastruttura per il *training* che metta a disposizione una piattaforma per l'erogazione diretta di moduli formativi (sia in modalità sincrona che asincrona) e una per il deposito dei materiali didattici e infine lo sviluppo di una metodologia comune per il design e l'adattamento di tali materiali.

Le tipologie di utenti oggetto della formazione sono state identificate in: ricercatori, esperti e all'inizio della loro carriera, con la necessità di integrare le proprie competenze allo scopo di gestire correttamente i dati della ricerca; tecnologi che hanno bisogno di sviluppare dataset e strumenti da aprire alla comunità secondo i dettami della Scienza Aperta; docenti e formatori di ambito universitario e professionale che possano integrare l'uso delle IR nei loro curricoli di insegnamento; studenti di corsi di laurea e di dottorato, che devono essere esposti il prima possibile alle buone pratiche della comunità scientifica; professionisti a supporto della ricerca come *data stewards* e *data curators* per cui è necessario trasmettere le proprie competenze a diversi tipi di utenti; e infine i professionisti delle istituzioni culturali (archivi, biblioteche, musei), che possono beneficiare dei servizi offerti dalle IR. Per sondare le necessità formative dei vari gruppi, il WP dedicato al training lavora in stretto contatto con il WP dedicato all'analisi delle comunità di riferimento, che ha sviluppato un questionario e lo ha somministrato in via preliminare a una selezione di utenti. Inoltre, sono state implementate misure per il coordinamento con le varie associazioni, con cui a breve saranno presi i primi contatti.

L'infrastruttura di *training* diventerà una componente fondamentale del Marketplace H2IOSC e sarà composta da due piattaforme separate: un Virtual Training Environment, cioè un *learning management system* per l'erogazione di corsi in modalità sincrona e asincrona che permetterà ad H2IOSC e alle IR coinvolte di offrire formazione diretta sia al proprio personale che alle comunità di utenti; e la piattaforma FAIR and Interoperable Training Materials Publication Platform allo scopo di conservare, condividere, citare e riutilizzare i materiali formativi come oggetti digitali FAIR dotati di licenza e di un sistema di versionamento, in modo da creare una comunità di formatori che condividano e riutilizzino i rispettivi materiali all'interno dei programmi accademici e di formazione. Questi requisiti tecnici e funzionali sono stati tradotti in

[4] https://www.go-fair.org/fair-principles/
[5] I materiali di *training* sviluppati da DARIAH-EU sono disponibili su https://campus.dariah.eu/. All'interno del framework, le collezioni esterne sono cercabili tramite lo strumento DARIAH Pathfinders https://campus.dariah.eu/source/dariah-pathfinders/page/1
[6] https://training-toolkit.sshopencloud.eu/entities?search=&f%5B0%5D=content_type%3Asource
[7] https://sshopencloud.eu/
[8] https://eosc-portal.eu/

documenti di specifica destinati a bandi pubblici per l'implementazione di entrambe le piattaforme, le cui procedure di aggiudicazione si sono appena concluse.

Quanto allo sviluppo di materiali di *training*, si è deciso di dedicare una prima fase all'adattamento di moduli didattici preesistenti, originati dalle iniziative di formazione già in atto nelle IR coinvolte, alle necessità degli utenti di H2IOSC. In un primo momento, l'erogazione di questi moduli sarà dedicata alla formazione del personale interno delle IR coinvolte in una prospettiva di potenziamento delle competenze, ma sono in programma proposte di disseminazione come, ad esempio, la partecipazione del nodo italiano di CLARIN alla *summer school* Digital Tools for Humanists[9] organizzata dall'Università di Pisa sotto la direzione del Laboratorio di Cultura Digitale. Successivamente, è previsto il potenziamento dell'offerta formativa di ciascuna IR con la creazione di nuovi corsi specifici per la propria comunità di riferimento e infine verranno messe in atto iniziative congiunte per consolidare le competenze trasversali del bacino di utenza H2IOSC e formazione specifica per i membri del consorzio affinché possano a loro volta diventare formatori esperti nelle buone pratiche della Scienza Aperta e dei principi FAIR. A coordinamento delle operazioni, è stata sviluppata una metodologia di adattamento e design dei materiali didattici come oggetti digitali FAIR che verrà dettagliata nella prossima sezione.

3. MATERIALI DIDATTICI FAIR: UNA METODOLOGIA CONDIVISA

Consapevoli che la produzione ed erogazione diretta di moduli didattici potrà soddisfare solo in parte le esigenze della comunità, H2IOSC vede nella costruzione di un ecosistema di FAIRificazione e condivisione di materiali didattici per il potenziamento e la trasmissione di competenze digitali una missione importante, anche al fine valorizzare le molte delle iniziative già messe in campo da AIUCD. Questa sezione illustra in particolare i principi alla base dello sviluppo della Publication Platform, dove andranno depositati dapprima tutti i materiali sviluppati dalle realtà coinvolte in H2IOSC e dai partner ed in seguito anche materiali della comunità scientifica, al fine di massimizzarne il riutilizzo. Secondo la strategia dettagliata nella sezione precedente, lo sviluppo dei moduli didattici prenderà le mosse da una prima fase di raccolta e sistematizzazione di materiali preesistenti (*slide deck*, infografiche, materiale audiovisivo, collezioni digitali ecc.). Ciò comporta l'ideazione di un flusso di lavoro capace sia di allineare i materiali con l'applicazione dei principi FAIR che di integrare moduli già concepiti come FAIR, ma spesso descritti secondo standard differenti, in formati diversi e con licenze tra loro incompatibili. Per questo, nella definizione di una metodologia condivisa sono state definite pratiche comuni per la descrizione dei materiali con metadati standard, l'applicazione di licenze adatte al riuso e l'attribuzione di identificativi univoci e persistenti (PID) che ne permettano la corretta citazione e attribuzione ai rispettivi autori.

L'unità minima modulare di riferimento è stata individuata nella singola lezione descritta in accordo con la definizione di *learning object* come un pacchetto di una lezione, un'attività e una valutazione con un singolo obiettivo di apprendimento e un risultato di apprendimento concreto adottata dal progetto Skills4EOSC [2]. In questo modo la lezione, accompagnata da opportuni metadati, può essere utilizzata come base per la creazione di contenuti didattici più complessi. I metadati per descrivere ogni singolo *learning object* sono stati individuati nel Minimal Metadata Set for Learning Resources proposto dalla Research Data Alliance [4] che stabilisce 14 campi suddivisi in 3 categorie di informazioni (*descriptive*, *access*, *educational*) ed è attualmente in fase di valutazione presso progetti come OpenPlato[10], SSHOC Training Discovery Toolkit e NI4OS Training Platform[11]. Si tratta di uno schema flessibile, aperto ad eventuali integrazioni a seconda delle necessità connesse, ad esempio, alle diverse prospettive di insegnamento formali, professionali e informali, ed è volto a massimizzare la cercabilità dei dati senza appesantire il sistema dei descrittori, garantendo conformità e riusabilità del materiale esistente. In una prospettiva FAIR dal punto di vista dei formatori, i metadati più importanti sono quelli relativi alle parole chiave, che possono connettersi a vocabolari condivisi specifici per il settore disciplinare e quindi rendere il materiale facilmente trovabile; alle licenze utilizzate, che sono alla base delle possibilità di accesso e riutilizzo; e infine i campi connessi a caratteristiche didattiche specifiche come *target group*, *learning outcome(s)* ed *expertise level*, che permettono di valutare la coerenza della lezione con le proprie finalità didattiche ed eventualmente pianificarne l'adattamento.

In questa fase del lavoro si è deciso inoltre di optare per la conversione del materiale preesistente nei formati più aperti disponibili, in modo da garantire la massima libertà e flessibilità di riutilizzo. Il formato non proprietario più versatile è sicuramente il plaintext, che verrà preferito laddove possibile, ad esempio rispetto a formati PDF (.ppt) o Moodle (.mbz). Per il design di nuovo materiale, che si tratti di creazione ex novo o di una rielaborazione di moduli preesistenti, inclusi

[9] https://digitaltools.labcd.unipi.it/
[10] https://openplato.eu/blocks/catalog/list.php
[11] https://training.ni4os.eu/

quelli convertiti in plaintext, si prefigura l'utilizzo di Markdown come linguaggio open-source estensibile, intuitivo e facilmente convertibile in HTML, JSON, XML, YAML e molti altri formati. Lo stesso principio sarà alla base della scelta delle licenze[12], che saranno più aperte possibile in considerazione delle regolamentazioni vigenti per la tipologia di dati in oggetto e della compatibilità delle licenze applicate al materiale pregresso. Un aspetto importante per facilitare il lavoro dei formatori e promuovere l'integrazione e il riutilizzo di *learning objects* è costituito dalle modalità di citazione dei materiali riutilizzati. A questo scopo, H2IOSC ha in progetto di adottare pratiche formalizzate e condivise per permettere la citazione e il riutilizzo dei propri materiali didattici, sia originali che derivati, considerati in tutto e per tutto come prodotti della ricerca. Fondamentale in questo senso sarà anche l'interscambio con le comunità scientifiche ed in particolare con AIUCD, da sempre sensibile a queste tematiche, per la definizione di modalità di citazione, riuso e deposito chiare e di facile implementazione.

4. PROSPETTIVE FUTURE E CONCLUSIONI

Il passo successivo comporterà l'applicazione di tale strategia ai corsi di formazione specifici per ogni IR coinvolta a partire dal materiale già presente, fra cui: il corso "Introduction to Language Data: Standards and Repositories" sviluppato da CLARIN nell'ambito del progetto UPSKILLS[13]; il corso "La lessicografia web based all'Opera del Vocabolario Italiano: metodi e strumenti per la ricerca filologica e linguistica", elaborato dal gruppo di lavoro di DARIAH-IT con tutorial sugli strumenti descritti, e i webinar[14] della serie didattica online "Current Topics in Heritage Science" che verranno adattati dal gruppo di lavoro di E-RIHS. In seguito, saranno sviluppati dei progetti pilota su servizi innovativi nell'ambito delle diverse IR: questi prototipi, sotto forma di piattaforme o *hub*, riuniranno servizi, flussi di lavoro e interfacce specifici per il dominio disciplinare di afferenza e saranno progettati in modo da essere scalabili ed estendibili con l'aggiunta di nuove risorse e dati, al fine di acquisire conoscenze e adeguare gli obiettivi man mano che l'implementazione procede. Alcuni tra i progetti pilota previsti sono, ad esempio i tutorial che saranno sviluppati da CLARIN per facilitare l'uso dei Linguistic Linked Open Data, per la cura e il deposito di dati di storia orale e per la curatela di dati neurolinguistici e psicolinguistici. Infine, vi saranno corsi dedicati all'uso dei servizi di H2IOSC come, ad esempio, il futuro Marketplace, e ad aspetti gestionali delle IR come l'Infrastructure Management per le Digital Humanities, con particolare focus sulla creazione di *data management plan* e aspetti legali ed etici connessi alla protezione dei dati digitali per le tipologie di dati rilevanti per H2IOSC.

Le potenzialità di una metodologia condivisa per il design e l'adattamento di materiali didattici come prodotti della ricerca sono molteplici, primi fra tutti il superamento di gap metodologici e terminologici esistenti tra ambiti disciplinari diversi e l'agevolazione della formazione professionale nell'ambito di progetti nazionali e internazionali. Per questo, in accordo con i principi della Scienza Aperta, il presente contributo apre alla comunità la metodologia sviluppata in seno al progetto H2IOSC per l'adattamento e il design di materiali didattici come *learning objects* modulari e riutilizzabili in modo da massimizzare l'impatto delle iniziative di formazione che integrano l'uso delle Infrastrutture di Ricerca.

5. RINGRAZIAMENTI

Progetto H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 4 "Istruzione e Ricerca" Componente 2 "Dalla ricerca all'impresa" Linea di Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione" Azione 3.1.1 "Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti" – Codice progetto IR0000029 - CUP B63C22000730005. Soggetto attuatore CNR.

BIBLIOGRAFIA

[1] Degl'Innocenti, Emiliano, Monica Monachini, Alberto Bucciero, Enrico Pasini, Bruno Fanini, e Francesca Frontini. «H2IOSC: Humanities and Heritage Open Science Cloud». In *La memoria digitale: forme del testo e organizzazione della conoscenza. XII Convegno Annuale AIUCD, Siena, 5-7/06/2023. Proceedings*, a cura di Emmanuela Carbé, Gabriele Lo Piccolo, Alessia Valenti, e Francesco Stella, 63-64, 2023. https://doi.org/10.6092/unibo/amsacta/7721.

[12] https://creativecommons.it/chapterIT/index.php/license-your-work/

[13] Il progetto UPSKILLS è stato un partenariato strategico Erasmus+ con lo scopo di identificare e affrontare le lacune nelle competenze degli studenti di linguistica attraverso l'integrazione dei curricola esistenti con materiali didattici di supporto: https://upskillsproject.eu/. Il materiale del corso è disponibile qui: https://upskillsproject.eu/project/standards_repositories/

[14] Si tratta di materiale originato da incontri mensili di formazione online su aspetti fondamentali dell'Heritage Science: https://www.iperionhs.eu/lectures-series/

[2] Filiposka, Sonja. «D2.2 Methodology for FAIR-by-Design Training Materials». Zenodo, 31 agosto 2023. https://doi.org/10.5281/ZENODO.8305540.

[3] Frontini, Francesca, e Monica Monachini. «Infrastrutture digitali per le scienze umane e sociali». In *Digital Humanities. Metodi, strumenti, saperi*, 197–213. Roma: Carocci, 2023.

[4] Hoebelheinrich, Nancy J., Katarzyna Biernacka, Michelle Brazas, Leyla Jael Castro, Nicola Fiore, Margareta Hellström, Emma Lazzeri, et al. «Recommendations for a Minimal Metadata Set to Aid Harmonised Discovery of Learning Resources». Zenodo, 9 giugno 2022. https://doi.org/10.15497/RDA00073.

Rethinking scholarly digital objects as cultural heritage: the KNOT project

Laurent Fintoni
Dipartimento di Filologia Classica e Italianistica, Università di Bologna, Italia – laurent.fintoni2@unibo.it

ABSTRACT
This paper presents the KNOT project, a three-year pilot tasked with investigating ways to integrate the digital cultural heritage of Italian universities within the national infrastructure being developed by the Ministry of Culture, and its central argument for rethinking the digital objects produced by academic research projects as interesting and, so far, unexplored examples of this digital cultural heritage. The paper discusses the key steps in the development of a conceptual framework for rethinking scholarly digital objects as cultural heritage starting with the definition of these objects and identification of the potential heritage values they hold, the selection of the humanities, and in particular the digital humanities, as the academic field from which to select these objects, the use of a census to evaluate the validity of these choices, and some of the issues that arose during this work around the classification, documentation, and visibility of these objects.

KEYWORDS
Digital Humanities; Digital Cultural Heritage; Italian Universities.

1. INTRODUCTION
Italian cultural heritage (CH) is diverse, omnipresent, and connected to the many histories found across its territory [4]. Among the institutions tasked with preserving and managing this heritage, universities stand out due to how their holdings are often intertwined across the three missions they must adhere to – teaching, research, and the dissemination of knowledge into wider society. However, while universities hold an important part of Italy's CH it has remained partially hidden within the national picture due to, in part, the late development of national discussions around its role and function which only began at the turn of the 21st century [12]. This situation has since been further complicated by the impact of digital technology, with universities having to contend with both analog holdings and newer digitized and born-digital objects that represent a digital cultural heritage (DCH) that has only grown in importance as the years have passed.

The challenge of how to integrate this particular segment of DCH at the national level has taken on new life recently within the National Plan for the Digitalisation of Cultural Heritage (NPD), which sets out the strategic vision and guidelines from the Ministry of Culture for the transformation of the country's CH over a five-year period (2022-2026) as part of the National Recovery and Resilience Plan[1]. Drafted by the Central Institute for the Digitalisation of Cultural Heritage – Digital Library (ICDP), a new body setup by the Ministry in 2020, the NPD seeks to create a digital ecosystem for CH with the ICDP guiding institutions and places of culture, such as universities, in this process[2].

The KNOT project is a three-year pilot that takes place within the context of this national effort, part of an agreement between the University of Bologna and the Ministry to establish a research infrastructure to support the ICDP's efforts with parallel and connected research initiatives in three departments (Department of Cultural Heritage - DBC, Department of Classical Philology and Italian Studies - FICLIT, and Department of Computer Science and Engineering - DISI) tasked with investigating ways to integrate the DCH of Italian universities into this new digital ecosystem of culture. The first challenge our project had to face was how to effectively represent this DCH at a national level considering that to date there has been little consensus within institutions or across them as to what constitutes it[3]. To this end, we decided to take a lateral approach to the problem and consider how universities are already in possession of a multifaceted and interesting, and so far unexplored, example of DCH in the form of digital objects created by academic research projects. During our first year of research we focused on the development of a conceptual framework for rethinking these scholarly digital objects as DCH, with the framework intended to act as a base for the primary expected outputs of the pilot: a web application (to act as an example of early adoption of the ICDP's infrastructure) and guidelines for the collection, management, and enrichment of this DCH at the national level. In this paper we present some of the key steps in the development of the conceptual framework including the definition of these objects and identification of the potential heritage values they hold, the selection of the humanities, and in particular the digital humanities (DH), as the academic field from which to select

[1] https://docs.italia.it/italia/icdp/icdp-pnd-docs/it/v1.1-febbraio-2023/index.html
[2] https://digitallibrary.cultura.gov.it/chi-siamo/
[3] A parallel line of research being conducted by the DBC has been tasked with conducting a census of the DCH of Italian universities and the remark about the lack of consensus stems from the results of their first year of investigation.

these objects, the use of a census to evaluate the validity of these choices, and, lastly, the issues that arose during this work around the classification, documentation, and visibility of these objects.

2. DEFINING SCHOLARLY DIGITAL CULTURAL HERITAGE

The definitions of DCH put forward by the United Nations Educational, Scientific and Cultural Organisation (UNESCO) at the international level and by the PND at the national level provide a starting point for our rethinking of digital scholarly objects as DCH. UNESCO adopted its Charter on the Preservation of Digital Heritage in October 2003, defining the scope of digital heritage as consisting of "unique resources of human knowledge and expression", whether digital-born or converted from analogue, stemming from a plurality of backgrounds (educational, legal, scientific, and others) and represented by a multiplicity of formats, from text to databases, software to web pages[4]. The PND meanwhile defines DCH in a more focused way, honing on "the set of digital objects produced by the modeling of data or by the organization of digitally-native content to achieve more advanced knowledge objectives through the development of the relational potential that characterizes its dissemination" which taken together and understood within an ecosystem logic contribute to the formation of a CH similar to the one assigned to tangible and intangible assets[5].

From these definitions as well as insights from the heritage sector that call to move our thinking past digital objects as data towards a form that contains multiple agencies and spatio-temporal distributions [2: 279] we can think of the scholarly digital objects we are interested in to therefore comprise not just collections of information (such as datasets) but, crucially, also digital forms that enable interaction with information, from software and data services such as search interfaces or APIs to visualization and annotation tools. Considering these objects thusly we can then begin to think past the scientific context in which they are most commonly understood, as well as evaluated, and look at what heritage value might be found in their "reservoir of meaning" [2: 34]: the activity that produced them; relationships to the information they encode; new contexts they create for this information; and the ways in which they can foster the acquisition of new knowledge from this combination of context and information.

Having narrowed our definition of DCH to specific objects of interest based on our argument, next we needed to narrow the scope of academic research to a specific field from which we could draw these objects. Considering that our research takes place within the Digital Humanities Advanced Research Center, a DH research center at the University of Bologna, the humanities presented itself as the ideal field of reference for the project, a choice that was strengthened by recent research and previous national efforts to engage with the DCH of universities.

At the national level, there have been a handful of humanities projects focused on the digitized holdings of universities, among them MICHAEL (Multilingual Inventory of Cultural Heritage in Europe) in the mid-2000s, a census of digitized collections from 77 Italian universities for inclusion in a European initiative coordinated by the University Library Center of the University of Padua[6], and POMUI (POrtal MUseums Italian) in the early 2010s, the country's first network of university museums in which 12 institutions participated in an inventory and electronic catalog of their holdings and the creation of a bilingual web portal to raise awareness of and promote this particular segment of their CH[7]. More recently, a bibliometric analysis of CH research in the humanities, based on data from the Web of Science platform (journal articles published between 2003 and 2022), showed how central the field has become to Italian academia's engagement with CH as a whole: Italian humanities scholars are among the top three in the world for the production of scholarly articles about CH and among the most cited [17]. Meanwhile, an analysis of projects presented at the yearly conference of the Italian Association for Digital Humanities and Culture (AIUCD) between 2018 and 2020 identified disciplines of the text (publishing, philology, literature, and linguistics) and the management of DCH as the dominant areas of focus for Italian digital humanists [14]. The DH in particular offer further opportunities for our project as they promote interdisciplinary approaches [15] and project-oriented practices where existing CH objects are often used as source material for the creation of new digital objects that seek to recontextualize this source material and promote engagement from end users. This not only reflects some of the heritage values that can be found in digital forms, as mentioned above, but also mirrors the goals of I.PaC (Infrastruttura e servizi digitali per il Patrimonio Culturale)[8], the online infrastructure the ICDP is building and within which the results of the KNOT project are expected to be included.

[4] https://unesdoc.unesco.org/ark:/48223/pf0000133171
[5] See glossary in https://docs.italia.it/italia/icdp/icdp-pnd-docs/it/v1.1-febbraio-2023/index.html
[6] https://bibliotecadigitale.cab.unipd.it/biblioteca-digitale/progetti/michael
[7] http://www.pomui.unimore.it/site/en/home/project-2012.html
[8] https://ipac.cultura.gov.it/

3. A CENSUS OF ITALIAN HUMANITIES PROJECTS

We conducted a census of existing Italian humanities research projects and analyzed the results to evaluate the validity of our choices as well as extract potential features and characteristics to use in the development of a data model based on our conceptual framework[9]. We began our census by looking at research projects listed on the AIUCD[10] and European Association of Digital Humanities (EADH)[11] websites. We then visited the websites of specific humanities departments, research centers, and laboratories that were included in these lists as well as those of members to look for additional relevant projects. Finally, we also conducted manual searches within the official websites of universities, including digital libraries, for further potential projects and lists of relevant resources. At each step we focused on recording projects, whether finished or ongoing, that had produced one or more digital objects that could be considered as representative of our argument such as catalogs, databases, software, corpora, archives, and scholarly editions. This produced a list of 91 projects for which we then recorded a variety of metadata based on available information including: links to websites and data repositories; the entities involved (universities, departments, labs, research centers); the year of creation; the primary subject and time period; the types of objects used and created; the tools and technologies used in the creation of digital objects; the research activities (based on TaDiRAH, a taxonomy of digital research activities in the humanities[12]); the licenses; the documentation; and the degrees to which projects made their results available via Linked Open Data practices or technologies such as APIs.

The results of the census presented us with an overall picture that featured a mix of projects that produced what we can consider as traditional humanities objects such as corpora, digital editions, catalogs, and databases alongside some examples of newer ones such as visualization tools, machine learning pipelines, software, and platforms that reflect some of the digital forms of interest mentioned in section 2. In terms of the disciplines involved in the projects, the prominence of philology, literature, and linguistics that [14] noted as key areas of interest for Italian digital humanists was also visible in our results alongside contributions from the fields of design, musicology, archeology, archival science, and computer science. While these results gave us confidence that the humanities could offer a variety of objects and approaches to use in our project, they also evidenced a number of issues for us to consider: the prominence of certain fields and objects, how to classify the scholarly digital objects we are interested in, the quality of available documentation and how necessary information can be extracted from it, and the visibility of these objects within the specifics of the Italian academic system.

4. ISSUES AND CHALLENGES

The first issue we had to consider was how many of the projects referred to the objects they created in their documentation using different, sometimes complimentary, sometimes contradictory, terms. For example, Archivio della Latinità Italiana del Medioevo (ALIM)[13] uses the term archive in its title and project description but the term digital library in the interface. As a scholarly activity ALIM is creating a digital archive, in that it preserves items because of their perceived value to one or more communities (in this case Latin texts produced in Italy during the Middle Ages), yet the digital form created by activity, the scholarly digital object through which this archive is accessed, more closely resembles a digital library with content classified into collections, available for retrieval, and with additional services for users. Another example is DanteSources[14], which refers to the object it produced in its documentation as a digital library, tool, and knowledge base with the latter two more closely reflecting the affordances of the object (especially when you consider that the interface only communicates the ability to search) while digital library feels less appropriate in that there are no practical resources to be found and retrieved via the object but rather information extracted from the analysis of resources. Based on the results of our census we found that typological terms such as archive, library, catalog, and database were most commonly treated as interchangeable in documentation and interfaces while digital edition was the most consistently, and accurately, used term.

As our work progressed it became apparent that we needed a way to classify scholarly digital objects produced by humanities research using a reliable, empirical system such as a taxonomy. However, while there have been efforts within the international DH community to create useful taxonomies of research methods, tools, and activities – notably the DiRT Directory of digital research tools [13], the TAPoR gateway for text analysis tools[15], and the aforementioned TaDiRAH

[9] Documentation of the data model is available at https://icdp-digital-library.github.io/KNOT/website/ENG/data_model.html
[10] http://www.aiucd.it/progetti/
[11] https://eadh.org/projects
[12] https://tadirah.info/pages/Browser.html
[13] http://alim.unisi.it/
[14] https://dantesources.dantenetwork.it/en/index.html
[15] https://tapor.ca/home

(the latter two integrating elements of or knowledge from DiRT [8, 1]) – less focus appears to have been given to the objects such methods, tools, and activities might produce.

Parallel to this issue of taxonomy we also noticed that many of the common technologies used in the creation of scholarly digital objects, such as programming languages like Python, formats and standards like JSON or TEI, and software packages like Omeka, were not always reliably included or detailed in existing authority controls and thesauri, whether well-established ones such as the Library of Congress Subject Headings and the Getty Art & Architecture Thesaurus or more specific ones such as the OntoPiA network maintained by the Ministry of Culture[16] or the DHA Taxonomy from the Austrian Centre for Digital Humanities and Cultural Heritage[17].

With this in mind, as part of the data model we are developing we have created two SKOS-based controlled vocabularies (one taxonomy and one thesaurus) aimed at answering these specific classification needs while also providing the necessary controls for the eventual integration of this information into a database and retrieval system [10][18]. Furthermore, these vocabularies are intended to act as a base for ongoing reflection regarding definitions and meanings. Taking a cue from the rethinking of methods and activities as scholarly primitives that motivated the creation of TaDiRAH [1, 16] we can see how thinking about the definition of a set of primitives for scholarly digital objects, reflecting the basic functions that these objects fulfill for scholarly activity in the humanities alongside their practical functions, which may push against accepted definitions, may be useful both within and outside of the field. Such insights and reflections will feed into the guidelines the KNOT project is expected to produce at the end of its duration.

Connected to this issue of classification is the inconsistent quality of the documentation of the scholarly digital objects recorded in our census. Inconsistent documentation makes it more difficult to accurately describe such objects using a data model by obscuring their narrative [5]. Previous studies of documentation of DH resources, going back to the late 2000s, have found that documentation is often insufficient or lacks clear visibility [18, 7, 6] and this is something we have also seen in our work to date with projects often either lacking or obscuring important information such as clear summaries or general descriptions, their technology stack, their status, and the licenses available for use of the objects created (whether data or service). One notable exception is the handful of projects we have recorded that openly strive to meet the FAIR principles for scientific data management and stewardship [19], which often results in clearer, more accurate and extensive documentation. Returning to the previous examples, both ALIM and DanteSources offer fairly complete documentation on their website though the latter does not detail its technology stack nor indicates the status of the project or licenses for use of its data and services. Conversely, LiLa: Linking Latin[19] is a project that strives to meet FAIR requirements and provides the most complete documentation of all the activities catalogued by our census.

A recent checklist for documentation of humanities data that builds on previous studies recommends 13 components of consideration [11] that match many of the same problematic areas we have encountered in the evaluation of our census results (including general information, scope, functionality, data provenance, data access and reuse, and publications), underlining the importance of the issue while also providing a useful point of reference for the guidelines that the KNOT project is expected to produce.

Lastly, our census also brought us to consider issues of visibility for the digital objects we are interested in. While Italy boasts an active DH community, there are no official positions for the discipline, as exist in other countries, leading to difficulties in both professional recognition for academics as well as for the products of their research [9]. This is further compounded by the ways in which the National University Council organizes disciplinary areas and how it and the National Agency for the Evaluation of the University and Research Systems evaluate scientific production which leaves interdisciplinary research, central to the DH as noted previously, subject to evaluation under a specific area (rather than across the different areas of the different disciplines involved) and many of the digital objects produced by such research as not always worthy of scientific recognition [15]. While the growth of European research infrastructures focused on the arts and humanities, such as DARIAH[20] and CLARIN[21], has helped in this regard by providing avenues for the dissemination of digital scholarly objects and collaboration at the international level, at the national level visibility for the scholarly digital objects we are interested in remains a critical issue. As we found out during our census, within the current Italian system these objects and their various components are often scattered across institutional and public repositories

[16] https://github.com/italia/daf-ontologie-vocabolari-controllati/wiki
[17] https://vocabs.acdh.oeaw.ac.at/dha_taxonomy/en/
[18] Where relevant we have made use of existing definitions from the LoC's Subject Headings, the Getty AAT, and the ACDH's DHA taxonomy to create semantic relationships in our vocabularies. The vocabularies can currently be accessed at https://github.com/icdp-digital-library/KNOT/tree/main/data_model
[19] https://lila-erc.eu/
[20] https://www.dariah.eu/
[21] https://www.clarin.eu/

and websites without clear connections. A common example of this is that a database produced by a research project will have an accessible public interface on an official site hosted on an institutional sub-domain, yet this official site may not be linked to from the primary institutional domain. In turn, publications related to the creation of the database may be included in an institutional repository, such as the many academic ones that run the IRIS solution offered by CINECA, without a clear link to the official site where the online interface can be accessed. Lastly, some of the components of the database itself, such as the underlying data, may be available via an external repository such as GitHub but not clearly connected to the official site. In the case of our research this has a direct impact on the amount of work necessary to both evaluate and describe a project and the objects it created, and it likely also has an impact on the visibility of these digital scholarly objects for other potential users. While it is beyond the scope of the KNOT project to address issues of visibility caused by the specifics of the Italian academic system, we think that arguing for digital scholarly objects to be considered as examples of DCH worthy of integration in the nascent national infrastructure has the potential to raise their visibility elsewhere and offer insights into how to address issues of access and connectivity between humanities research projects and the objects they produce.

5. CONCLUSION AND FUTURE WORK

The development of a national infrastructure for DCH such as the one currently being undertaken by the ICDP offers a unique opportunity for Italian universities to engage with this segment of their CH holdings in new ways. To do this will require a broadening of the understanding of DCH that takes into consideration how digital objects can be more than data and how meaningful heritage value lies in digital forms that enable interaction with information. By focusing on the scholarly digital objects produced by humanities research, and in particular DH, as relevant examples of such DCH already held by universities, the KNOT project hopes to provide a useful base from which to let new insights emerge around how to valorize these objects both within and outside of academia.

The challenges we have highlighted around classification, documentation, and visibility have been key areas of focus in the development of the KNOT data model and its integration, alongside the results of the census, within an online catalog intended to form a first step towards an eventual web application. The first version of the data model and the online catalog[22] were made public in late 2023 and early 2024 respectively, including the controlled vocabularies. Updates to both the data model and controlled vocabularies are planned throughout the duration of the pilot, with a first round of updates in the spring of 2024 to reflect insights from their usage in the catalog. Further work includes: alignment with the ICDP for ingestion of the data produced by KNOT into the I.PaC, expected to begin in spring of 2024; the publication of the controlled vocabularies through an open-source browser such as Skosmos[23], which would fulfill all the requirements of making the vocabularies FAIR based on the rules outlined in [3]; engagement with the academic community to assess the usefulness of the data model; and further development of a web application and guidelines.

REFERENCES

[1] Borek, Luise, Quinn Dombrowski, Jody Perkins, and Christof Schöch. 'Scholarly Primitives Revisited: Towards A Practical Taxonomy of Digital Humanities Research Activities and Objects'. Zenodo, 2014. https://doi.org/10.5281/ZENODO.10866.

[2] Cameron, Fiona. The Future of Digital Data, Heritage and Curation in a More-than Human World. London; New York: Routledge/Taylor & Francis Group, 2021.

[3] Cox, Simon J. D., Alejandra N. Gonzalez-Beltran, Barbara Magagna, and Maria-Cristina Marinescu. 'Ten Simple Rules for Making a Vocabulary FAIR'. *PLoS Computational* 17, no. 6 (2021). https://doi.org/10.1371/journal.pcbi.1009041.

[4] Donato, Fabio, and Enrica Gilli. 'Un approccio "multi-scala" per la gestione del patrimonio culturale italiano'. *IL CAPITALE CULTURALE. Studies on the Value of Cultural Heritage*, 2011, 197–225. https://doi.org/10.13138/2039-2362/118.

[5] Fafinski, Mateusz. 'Facsimile Narratives: Researching the Past in the Age of Digital Reproduction'. *Digital Scholarship in the Humanities* 37, no. 1 (2022): 94–108. https://doi.org/10.1093/llc/fqab017.

[6] Franzini, Greta, Melissa Terras, and Simon Mahony. 'Digital Editions of Text: Surveying User Requirements in the Digital Humanities'. *Journal on Computing and Cultural Heritage (JOCCH)* 12, no. 1 (2019): 1–23. https://doi.org/10.1145/3230671.

[7] Gibbs, Fred, and Trevor J. Owens. 'Building Better Digital Humanities Tools: Toward Broader Audiences and User-Centered Designs'. *Digital Humanities Quarterly* 6, no. 2 (2012). http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html.

[8] Grant, Kaitlyn, Quinn Dombrowski, Kamal Ranaweera, Omar Rodriguez-Arenas, Stéfan Sinclair, and Geoffrey Rockwell. 'Absorbing DiRT: Tool Directories in the Digital Age'. *Digital Studies* 10, no. 1 (2020). https://doi.org/10.16995/dscn.325.

[9] Hall, Crystal. 'Digital Humanities and Italian Studies: Intersections and Oppositions'. *Italian Culture. Informa UK Limited* 37, no. 2 (2019): 97–115. https://doi.org/10.1080/01614622.2019.1717754.

[22] https://projects.dharc.unibo.it/knot/

[23] https://skosmos.org/

[10] Harpring, Patricia. Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works. I. Los Angeles: Getty Research Institute, 2010.

[11] Middle, Sarah. 'A Documentation Checklist for (Linked) Humanities Data.' *International Journal of Digital Humanities*5, no. 2 (2023): 353–71. https://doi.org/10.1007/s42803-023-00072-z.

[12] Mozzoni, Isabella, Simone Fanelli, and Chiara C. Donelli. 'Italian University Collections: Managing the Artistic Heritage of the University's Ivory Tower'. *Uropean Journal of Cultural Management and Policy* 8 (2018): 31–43.

[13] Perkins, Jody, Quinn Dombrowski, Borek Borek, and Christof Schöch. 'Project Report: Building Bridges to the Future of a Distributed Network: From DiRT Categories to TaDiRAH, a Methods Taxonomy for Digital Humanities'. In *Proceedings of the 2014 International Conference on Dublin Core and Metadata Applications*, edited by William Moen and Amy Rushing, 181–83. Austin, Texas: Dublin Core Metadata Initiative, 2014.

[14] Salvatori, Enrica. 'Digital Public History Inside and Outside the Box'. *Magazén* 2 (2020). https://doi.org/10.30687/mag/2724-3923/2020/02/003.

[15] Tammaro, Anna M. 'Evaluation of Digital Humanities: An Interdisciplinary Approach'. In *Bridging Between Cultural Heritage Institutions*, edited by Tiziana Catarci, Nicola Ferro, and Antonella Poggi, 136–46. Communications in Computer and Information Science. Berlin, Heidelberg: Springer, 2014. https://doi.org/10.1007/978-3-642-54347-0_15.

[16] Unsworth, John. 'Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?' *Symposium on Humanities Computing: Formal Methods, Experimental Practice. King's College, London* 13 (2000). https://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

[17] Vlase, Ionela, and Tuuli Lähdesmäki. 'A Bibliometric Analysis of Cultural Heritage Research in the Humanities: The Web of Science as a Tool of Knowledge Management'. *Humanities and Social Sciences Communications* 10, no. 1 (2023): 1–14. https://doi.org/10.1057/s41599-023-01582-5.

[18] Warwick, Claire. 'If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data'. *Literary and Linguistic Computing* 23, no. 1 (2007): 85–102. https://doi.org/10.1093/llc/fqm045.

[19] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3, no. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

The ATLAS: a knowledge graph of digital scholarly research on Italian Cultural Heritage

Marilena Daquino[1], Alessia Bardi[2], Marina Buzzoni[3], Riccardo Del Gratta[4], Angelo Mario Del Grosso[5], Franz Fischer[6], Francesca Tomasi[7], Roberto Rosselli Del Turco[8]
[1]Università di Bologna, Italia - marilena.daquino2@unibo.it
[2]CNR-Istituto di Scienza e Tecnologie dell'Informazione, Italia - alessia.bardi@isti.cnr.it
[3]Università Ca' Foscari Venezia, Italia - mbuzzoni@unive.it
[4]CNR Istituto di Linguistica Computazionale, Italia - riccardo.delgratta@ilc.cnr.it
[5]CNR Istituto di Linguistica Computazionale, Italia - angelo.delgrosso@ilc.cnr.it
[6]Università Ca' Foscari Venezia, Italia - franz.fischer@unive.it
[7]Università di Bologna, Italia - francesca.tomasi@unibo.it
[8]Università di Torino, Italia - roberto.rossellidelturco@unito.it

ABSTRACT

ATLAS is a research initiative that aims to improve the FAIRness and exploitation of Digital Humanities (DH) projects and scholarly data about Italian Cultural Heritage (CH). This contribution describes the main challenges and opportunities of DH projects related to discoverability, interoperability, and preservation. It also explains the methodology and objectives of ATLAS, which involves the integration and reengineering of metadata from selected sources and software solutions (referred to as pilots within the ATLAS context) into a knowledge graph using Semantic Web and Natural Language Processing technologies.

The expected outcomes and impacts of ATLAS, are (i) the definition of guidelines and best practices for DH projects; (ii) the creation of a reference set of excellence initiatives; (iii) the reconciliation of data with authority records and open data sources; (iv) the publication and preservation of the knowledge graph; and (v) the development of a platform for exploration and discovery of DH projects and resources, in synergy with the European Research Infrastructures CLARIN and OpenAIRE.

KEYWORDS

Digital Humanities; Knowledge Graph; Semantic Web; Research Infrastructures; Italian Cultural Heritage.

1. INTRODUCTION

ATLAS is a project funded by the Next Generation program of the European Commission for 24 months (October 2023 - October 2025) whose main goal is improving FAIRness (Findable, Accessible, Interoperable, Re-usable) [11] and exploitation of Digital Humanities (DH) projects and scholarly data about Italian cultural heritage[1].

Digital Humanities scholarly projects often promote the creation of Web-based resources wherein to collect authoritative data related to our heritage. However, web resources are often not easy to discover, and they risk obsolescence if not well documented and based on shared guidelines and standards. Moreover, projects are often self-referential, that is, they may not follow metadata standards and best practices. In addition, users' experience is limited in exploration, since interlinking across projects with a clear content overlap, and explanations of such overlaps - including contradictory statements or disagreement – are missing.

In this project, the potential deriving from having shared metadata standards, protocols, reusable workflows, good practices, guidelines, and evaluation frameworks in the Humanities will be investigated. To tackle the aforementioned problems in real-world scenarios, metadata about a pool of selected data sources and software solutions will be integrated into a knowledge graph and re-engineered with state-of-the-art Semantic Web technologies and Natural Language Processing approaches.

One of the objectives of the project is to define guidelines for excellence projects and to create a golden set of reference initiatives in the Digital Humanities. Sources will be selected on the basis of a number of criteria. Namely, sources should be: (1) published according to shareable criteria; (2) easily mined to extract research topics, inter and intra-textual relations between sources, as well as bibliographic, literary, and thematic data; (3) characterized by strategies for long-term preservation.

Moreover, data extracted will be reconciled with international authority records (e.g. VIAF) and open data sources (e.g. Wikidata) to facilitate reuse and the development of mashup applications. The produced knowledge graph will be published

[1] https://dh-atlas.github.io

according to the aforementioned guidelines and preserved by the ILC4CLARIN repository[2] of the Italian node of the National Consortium of the CLARIN Research Infrastructure, CLARIN-IT (https://www.clarin-it.it/it). The knowledge graph will be made available in a format compliant with international guidelines, to ensure compatibility with OpenAIRE[3] and with the European Open Science Cloud[4].
Finally, enhanced data will be leveraged in a dedicated platform to support the exploration and discovery of the landscape of DH projects, and will provide suggestions on tools and resources to scholars who are planning new projects. Thanks to the synergy with OpenAIRE, ATLAS will build on top of existing services to support discovery of research data beyond those analysed and integrated into the ATLAS knowledge graph.

2. STATE OF THE ART

In recent years, Italian cultural institutions have devoted their activities to digitising and serving interoperable Linked Open Data (LOD) – e.g. the Italian archival system[5], the Italian Ministry of Cultural Heritage[6] [2], Culturaitalia[7], and Europeana[8]. The long tail of Digital Humanities academic projects related to Italian Cultural Heritage, and possibly relying on these data, is instead struggling to find a research framework that shares domain-dependent best practices and fosters findability and reusability of research data.
Nowadays, several catalogues of DH research address scholarly digital editions[9] [5], lists of projects gathered by national associations (AIUCD), research centres (e.g. /DH.arc, VeDPH, DH@FBK), international associations (EADH), and disciplinary surveys[10] [7].
To the best of our knowledge, there are currently no representative, comprehensive, and reusable catalogues of scholarly data focused on DH projects based on the Italian Cultural Heritage, leveraging Semantic Web technologies for interlinking and dissemination.
Secondly, a clear definition of guidelines for creating DH projects is needed. Information quality and technical requirements exist, such as MESA and NINES, Monasterium.net, as well as guidelines such as RIDE, MDR, and Reviews in Digital Humanities, or best practices for FAIR data [8]. These standards and models are fundamental to the community and need to be extended to cover the variety of resources that populate the DH domain.
Another important aspect is the use of ontologies as a means for enriching the semantic expressivity of data. Existing ontologies, such as CIDOC-Conceptual Reference Model[11], SPAR ontologies[12] [10], HiCO[13] [4], and representation models, e.g. Nanopublication [6], provide a solid basis to many projects. Nonetheless, projects may adopt them in different ways, with different granularities, and therefore hinder the semantic interoperability of data.
Existing Natural Language Processing technologies [9] can be adapted to extract content from the full-text of resources and can contribute to harmonise data granularity and to foster interlinking between sources. To this extent, the idea of "knowledge graph" as a network of interlinked semantic data (proposed by Google in 2012) is a reference point for our work. Furthermore, research infrastructures and clusters (CLARIN, DARIAH, PARTHENOS, OpenAIRE, OpenCitations, TRIPLE, ARIADNE) offer methods to aggregate, preserve, and provide tools and resources in either IaaS (Infrastructure-as-a-service) or PaaS (platform-as-a-service) architectures to support researchers. The research infrastructures represent another important point to deal with, to foster collaboration (in particular with CLARIN and OpenAIRE) and to design the technological environment.

[2] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/
[3] http://www.openaire.eu
[4] https://eosc-portal.eu/
[5] https://san.beniculturali.it
[6] http://dati.beniculturali.it
[7] https://dati.culturaitalia.it
[8] https://data.europeana.eu
[9] A catalog of: Digital Scholarly Editions, v 4.0, 2020ff under the direction of Patrick Sahle, with Georg Vogeler (contributions 1998ff), Jana Klinger (catalog entries 2019ff), Stephan Makowski (technical system 2020ff) and Nadine Sutor (various support 2020ff). https://www.digitale-edition.de.
[10] Vuozzo, Alessandro, "Griseldaonline. Gli strumenti dell'Italianistica digitale. Una sitografia", 2020. https://site.unibo.it/griseldaonline/it/strumenti/strumenti-italianistica-digitale
[11] https://cidoc-crm.org
[12] http://www.sparontologies.net/
[13] http://purl.org/emmedi/hico

3. OBJECTIVES AND APPROACH TO RESEARCH

The goal of the ATLAS is to create a knowledge graph of the international scholarly research on Italian Cultural Heritage leveraged in a Web portal that fosters users' interpretation and engagement. The ATLAS aims to collect, enrich, and provide information extracted from online resources produced in the international academic context, such as digital textual corpora and editions (mostly based on the TEI/XML standard) as well as datasets following Linked Open Data principles and databases, but also ontologies and tools adopted in the community. The ATLAS leverages existing digital heritage resources to create an innovative system that allows for the reuse of scholarly projects in new, non-native environments. In light of the state of the art, the ATLAS has two research lines:

(1) Metadata identification from pilots and guidelines. Data from a selection of digital scholarly resources (text collections, digital editions, linked open datasets, ontologies, and tools for interacting with texts) is analysed. Namely, the following excellence Italian DH projects and resources have been selected:
Text collections:
- ALIM, the Archive of the Italian Latinity of the Middle Ages[14]
- Biblioteca italiana[15]
- BUP - Digital Humanities[16]

Digital Editions:
- VaSto, VArchi STOria fiorentina[17]
- Codice Pelavicino Digitale[18]
- Leges Langobardorum[19]
- Digital Edition of Aldo Moro's works[20]

Software:
- EVT, Edition Visualization Technology[21]
- Voyant tools[22]

Linked Open Datasets:
- Zeri & LODE [5]: the Linked Open Dataset of the Federico Zeri's photo archive[23]
- LiLa - Linking Latin[24]
- Biflow - Toscana Bilingue Catalogue[25]

Ontologies:
- CIDOC-CRM
- SPAR
- HiCO

The metadata analysis will serve for the definition of the ontology of the ATLAS knowledge graph and for the definition of a set of guidelines targeting DH projects.

(2) ATLAS Platform. Methods to extract data from sources, create, enrich, and access the knowledge graph - including a Web application integrating tools that can be applied to collected sources - will be designed and developed. Use cases built on data from the selected pilot projects will support the technical specification of the ATLAS platform. To support data discovery among a wider range of resources, beyond the data of the selected pilot projects, the ATLAS will exploit the OpenAIRE CONNECT Gateway on Digital Humanities and Cultural Heritage[26] [1] and will identify the subset related to Italian CH. The ATLAS can therefore benefit also from the full-text and data mining processes implemented by OpenAIRE to identify links between research data, literature, and funding projects. Moreover, the ATLAS Platform will feed OpenAIRE with its high quality metadata records and, consequently, will ensure the visibility of the knowledge graph in the context of the European Open Science

[14] http://en.alim.unisi.it/
[15] http://www.bibliotecaitaliana.it/
[16] https://bup.unibas.it/library/DH
[17] https://dharc-org.github.io/progetto-vasto/
[18] https://pelavicino.labcd.unipi.it/
[19] http://alim.unisi.it/editto-di-rotari/
[20] https://aldomorodigitale.unibo.it/
[21] http://evt.labcd.unipi.it/
[22] https://voyant-tools.org/
[23] http://data.fondazionezeri.unibo.it
[24] https://lila-erc.eu/
[25] https://biflow.hypotheses.org/
[26] https://dh-ch.openaire.eu

Cloud (EOSC). The preservation of the knowledge graph will be guaranteed by the ILC4CLARIN repository operated by the Italian node of CLARIN.

4. EXPECTED OUTCOMES AND EVALUATION

The ATLAS will contribute to the Italian DH research community with four main results:

- A whitebook including results of the analysis of the state of the art, good practices for FAIR scholarly data and to ensure high-quality content.
- A knowledge graph on DH projects and scholarly data on Italian Cultural Heritage, accessible online via the ATLAS web application and preserved in ILC4CLARIN.
- The pilots evaluation, highlighting differences and strategies to cope with mapping knowledge, data manipulation, access and persistence of different types of digital artifacts.
- A search portal dedicated to scholarly literature and data relevant to the pilots and beyond, built on top of the OpenAIRE CONNECT Gateway on Digital Humanities and Cultural Heritage.

The ATLAS Web platform will advance the state of the art of DH research by fostering FAIRness of humanities research data, increasing their findability, and reusability, while allowing scholars, teachers, and students to perform data-driven research on high-quality scholarly data. New digital resources will be admitted into the platform on the basis of a peer review process. Criteria will be formalised in guidelines, protocols, and tutorials.

The knowledge graph, the services, and the web portal will be iteratively co-designed and evaluated on the basis of standard methodologies and on the continuous feedback provided by domain experts. Results will be documented and shared with the broad community of Humanities and Digital Humanities scholars and practitioners.

The aforementioned pilots will contribute to assess the development of the platform through continuous validation of its technologies. Pilots contribute to evaluate the following activities:

- quality and completeness of data retrieved via data harvesting, parsing and mining
- adequacy of guidelines, validated by adopters and stakeholders in the Italian DH community
- correctness of ontology mapping from specific domains to the project ontologies

The goal is to maximise synergy and methodological soundness. In particular, out of all pilots, for each category one project will be selected to validate results obtained on similar resources. Both quantitative metrics derived from standard evaluation frameworks (e.g. ontology consistency checks with reasoners, sampling and manual validation of text corpora) and qualitative metrics (e.g. user satisfaction, usability) will be designed to evaluate the three aforementioned points.

A first result of the project is the customization of CLEF 2.0, the new version of the CLEF (Crowdsourcing Linked Entities via web Form) application[27], in order to explore the potentiality of a tool to manage the creation of the LOD catalogue for the ATLAS project[28].

5. CONCLUSIONS AND FUTURE STEPS

We have presented the aims, methods and expected outcomes of ATLAS, a project funded by the Next Generation program of the European Commission, which seeks to enhance the FAIRness and exploitation of Digital Humanities (DH) projects and scholarly data on Italian Cultural Heritage. We have discussed the main challenges and opportunities that DH projects face in terms of research data management, long-term accessibility and exploitation. The project is in its first phase and it is analysing the selected excellence projects. Such analysis will support the definition of best practices, guidelines and the ontology of the knowledge graph. The knowledge graph is one of the main deliverables of the project and it will have a twofold impact: it will improve the visibility and re-usability of the selected excellence projects and their high-quality data, and it will foster their interoperability with major initiatives in the DH domain (CLARIN) and open scholarly communication (OpenAIRE).

The communication and dissemination activities in the research community of DH projects on Italian Cultural Heritage will be crucial to raise awareness of the benefits of adopting standards and best practices in terms of visibility, FAIRness, long-term preservation and sustainability. Moreover, they will also facilitate the expansion of the knowledge graph with information on additional projects and research data. We hope that this project will contribute to the advancement of knowledge and innovation in the field of DH and Italian Cultural Heritage, and that it will inspire further research and collaboration in this area, in synergy with other National and International initiatives.

[27] https://polifonia-project.github.io/clef/

[28] The CLEF 2.0 application will be presented at the same AIUCD2024 conference with the title: *CLEF 2.0. Soluzioni per la catalogazione nativa Linked Data del patrimonio digitale culturale italiano*.

6. ACKNOWLEDGEMENTS

This work is partially funded by the European Union - Next Generation EU.

REFERENCES

[1] Baglioni, Miriam, Alessia Bardi, Argiro Kokogiannaki, Paolo Manghi, Katerina Iatropoulou, Pedro Príncipe, André Vieira, et al. 'The OpenAIRE Research Community Dashboard: On Blending Scientific Workflows and Scientific Publishing'. In Digital Libraries for Open Knowledge, edited by Antoine Doucet, Antoine Isaac, Koraljka Golub, Trond Aalberg, and Adam Jatowt, 11799:59–69. Lecture Notes in Computer Science. Springer International Publishing, 2019. https://doi.org/10.1007/978-3- 030-30760-8_5.

[2] Carriero, Valentina A., Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 'ArCo: The Italian Cultural Heritage Knowledge Graph'. In *The Semantic Web - {ISWC} 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part {II}}*, edited by Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, 36–52. Lecture Notes in Computer Science. Auckland, New Zealand, 2019. https://doi.org/10.1007/978-3-030-30796-7_3.

[3] Daquino, Marilena, Francesca Mambelli, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. 'Enhancing Semantic Expressivity in the Cultural Heritage Domain'. *Journal on Computing and Cultural Heritage. Association for Computing Machinery (ACM)*, 31 July 2017, 1–21. https://doi.org/10.1145/3051487.

[4] Daquino, Marilena, and Francesca Tomasi. 'Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects'. In *Metadata and Semantics Research. MTSR 2015. Communications in Computer and Information Science*, edited by Emmanouel Garaoufallou, Richard J. Hartley, and Panorea Gaitanou, 544:424–36. Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24129-6_37.

[5] Franzini, Greta, Simon Mahony, and Melissa Terras. 'A Catalogue of Digital Editions'. In *Digital Scholarly Editing: Theories and Practices*, edited by Elena Pierazzo and Matthew Driscoll, 161–82. Cambridge (UK): Open Book Publishers, 2016. https://dx.doi.org/10.11647/OBP.0095.09.

[6] Groth, Paul, Andrew Gibson, and Jan Velterop. 'The Anatomy of a Nanopublication'. *Information & Use. IOS Press* 30, no. 1–2 (21 September 2010): 51–56.

[7] Hall, Crystal. 'Digital Humanities and Italian Studies: Intersections and Oppositions'. *Italian Culture. Informa UK Limited* 37, no. 2 (3 July 2019): 97–115. https://doi.org/10.1080/01614622.2019.1717754.

[8] O'Donnell, Daniel P., and Roberto Rosselli Del Turco. 'Good Things Come in Small Packages: Designing Distributed Editions and Tools for the Age of FAIR Data'. In *Presented at Assemblée Générale Du Consortium Cahier 2020, France, 27 November 2020*, 2020. https://doi.org/10.5281/ZENODO.4293723.

[9] Palmero Aprosio, Alessio, and Giovanni Moretti. 'Tint 2.0: An All-Inclusive Suite for NLP in Italian'. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-It. 10-12 Dicembre2018*, edited by Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, 311–17. Torino: Accademia University Press, 2018. https://doi.org/10.4000/books.aaccademia.3571.

[10] Peroni, Silvio, and David Shotton. 'The SPAR Ontologies'. In *The Semantic Web – ISWC 2018.*, edited by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, 11137:119–36. Lecture Notes in Computer Science. Springer International Publishing, 2018. https://doi.org/10.1007/978-3-030-00668-6_8.

[11] Wilkinson, Mark, Michel Dumontier, Ijsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3, no. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

INDICE DEGLI AUTORI