



# LA MEMORIA DIGITALE

FORME DEL TESTO E ORGANIZZAZIONE DELLA CONOSCENZA  
ATTI DEL XII CONVEGNO ANNUALE AIUCD

UNIVERSITÀ DI SIENA, 5-7 GIUGNO 2023

ASSOCIAZIONE PER  
L'INFORMATICA UMANISTICA  
E LA CULTURA DIGITALE



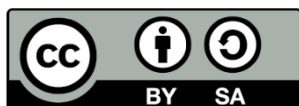
ISBN 978-88-942535-7-3

Copyright ©2023 AIUCD  
Associazione per l'Informatica Umanistica e la Cultura Digitale



Il presente volume e tutti i contributi sono rilasciati sotto licenza  
Creative Commons Attribution Share-Alike 4.0 International license ([CC-BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)).  
Ogni altro diritto rimane in capo ai singoli autori.

This volume and all contributions are released under the  
Creative Commons Attribution Share-Alike 4.0 International license ([CC-BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)).  
All other rights retained by the legal owners.



A cura di: Carbé, Emmanuela ; Lo Piccolo, Gabriele ; Valenti, Alessia; Stella, Francesco (2023). La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD, Siena: Università degli Studi di Siena

Ogni link citato era attivo al 22 maggio 2023, salvo ove diversamentee indicato.  
All links have been visited on 22th May 2023, unless otherwise indicated

Si prega di notificare all'editore ogni omissione o errore si riscontri, al fine di provvedere alla rettifica.  
Please notify the publisher of any omissions or errors found, in order to rectify them.

**aiucd.segreteria [at] aiucd.org**



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 Scientifico di AIUCD 2023.

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 2023 Scientific Committee.

Il programma della conferenza AIUCD 2023 è disponibile online

The AIUCD 2023 Conference Program is available online

<http://www.aiucd2023.unisi.it>

### **Comitato Scientifico**

Nicola Barbuti  
Marina Buzzoni  
Emmanuela Carbé (co-chair)  
Fabio Ciraci  
Fabio Ciotti  
Angelo Mario Del Grosso  
Maurizio Lana  
Monica Monachini  
Paolo Monella  
Roberto Rosseelli Del Turco  
Gino Roncaglia  
Francesco Stella (chair)  
Francesca Tomasi

### **Comitato organizzativo**

Elisabetta Bartoli  
Paola Bellomi  
Monica Bianchini  
Silvia Calamai  
Riccardo Castellana  
Rosalba Nodari  
Antonio Rizzo  
Enrico Zanini

**Segreteria del Convegno:** Gabriele Lo Piccolo, Francesca Pietrini;

Giulia Bassi, Silvia Cappa, Chiara Cauzzi, Martina Corti, Elena Crocicchia, Anna Guadagnoli, Giada Giannetti, Bogdan Groza, Alessia Luvisotto, Paola Mocella, Pietro Orlandi, Martina Paccara, Elisa Petri, Maria Grazia Schiaroli

Si ringrazia l'Ufficio stampa, comunicazione istituzionale e stampa digitale, l'Ufficio Ricerca, Biblioteche, Internazionalizzazione e Terza Missione e il Supporto eventi culturali e convegnistici dell'Università di Siena

Supporto tecnico: MCM Service, Santa Chiara Lab, Presidio San Niccolò

### **Enti organizzatori**

AIUCD;

Università degli Studi di Siena: Dipartimento di Filologia e critica delle letterature antiche e moderne (DFCLAM), Centro interuniversitario di Studi Comparati I Deug-Su, Centro Interdipartimentale di Ricerca Franco Fortini in "Storia della tradizione culturale del Novecento", Santa Chiara Lab, in collaborazione con i Dipartimenti di Ingegneria dell'Informazione e Scienze Matematiche (DIISM), di Scienze sociali, politiche e cognitive (DISPOC) e di Scienze storiche e dei beni culturali (DSSBC);

CLARIN-IT.

**Con il patrocinio di:** Journal of the Text Encoding Initiative

## **Chair di track**

### **Archivi, edizioni digitali, organizzazione della conoscenza**

Marina Buzzoni, Paolo Monella, Roberto Rosselli Del Turco

### **Analisi computazionale dei testi**

Fabio Ciotti, Rachele Sprugnoli

### **Intelligenza Artificiale e modelli applicati ai beni culturali**

Monica Bianchini, Federico Boschetti

### **Preservazione della memoria e del patrimonio digitale**

Nicola Barbuti, Maurizio Lana

### **Workshop**

Francesco Stella, Emmanuela Carbé

## **Lista dei revisori**

Stefano Allegrezza, Cristiano Amendola, Paolo Andreini, Laura Antonietti, Luca Bandirali, Sofia Baroncini, Elisabetta Bartoli, Stefano Bazzaco, Andrea Bellandi, Paola Bellomi, Benedetta Bessi, Andrea Bolioli, Luca Bombardieri, Simone Bonechi, Alice Borgna, Flavia Bruni, Paolo Buono, Dino Buzzetti †, Silvia Calamai, Anna Cappellotto, Giuliana Capriolo, Vittore Casarosa, Riccardo Castellana, Paola Castellucci, Simona Chiodo, Fabio Ciraci, Elisa Corrà, Elisa Cugliana, Fabio Cusimano, Christian D'Agata, Elisa D'Argenio, Vincenza D'Urso, Stefano Dall'Aglio, Marilena Daquino, Angelo Mario Del Grosso, Antonio Di Silvestro, Diego Mantoan, Dominique Brunato, Dominique Longrée, Edmondo Grassi, Elena Spadini, Giulia Fabbris, Pierluigi Feliciati, Paolo Fioretti, Franz Fischer, Greta Franzini, Francesca Frontini, Daniele Fusi, Simone Giusti, Marco Grasso, Fabiana Guernaccini, Alessandro Iannella, Benedetta Iavarone, Alessandro Lenci, Eleonora Litta, Agnese Macchiarelli, Marco Maggini, Elisabetta Magnanti, Francesco Mambrini, Tiziana Mancinelli, Anna Maria Marras, Cristina Marras, Luca Martinelli, Stefano Melacci, Federico Meschini, Alessio Miaschi, Andrea Micheletti, Giovanni Morrone, Rosalba Nodari, Giuseppe Palazzolo, Niccolò Pancino, Fiammetta Papi, Enrico Pasini, Marco Passarotti, Giulia Pedonese, Igor Pizzirusso, Federico Ponchio, Francesca Pratesi, Alessia Lucia Prete, Simone Reborà, Giulia Renda, Gino Roncaglia, Irene Russo, Enrica Salvatori, Eva Sassolini, Daniele Silvi, Daria Spampinato, Linda Spinazze', Francesco Stella, Matteo Tiezzi, Francesca Tomasi, Sara Tonelli, Gennaro Vessio, Paul Gabriele Weston, Michelangelo Zaccarello, Patrizia Zambrano, Marco Zappatore, Andrea Zugarini

**LA MEMORIA DIGITALE**  
**XII CONVEGNO ANNUALE AIUCD**  
**Siena, 5-7 giugno 2023**

**Indice**

<i>Prefazione</i>	IV
<i>In ricordo di Dino</i>	VII
<b>1. Archivi, edizioni digitali, organizzazione della conoscenza</b>	<b>1</b>
Creare e interrogare i modelli: un duplice valore conoscitivo. Il caso dei pareri di lettura einaudiani di Natalia Ginzburg	2
Exploring iconographical and iconological content in semantic data through art historians' interpretations	9
L'applicazione del riconoscimento testi neurale per la realizzazione di ristampe digitali	15
PATH, un thesaurus per le arti performative. Osservazioni empiriche e criticità d'uso	24
I dati della ricerca: gestione, diffusione e preservazione nei Next generation repositories	28
ODI and BACODI: a study on <i>Destini incrociati</i> by Italo Calvino with Semantic Web Technologies	35
L'edizione di un graphic novel medievale: il <i>Romanzo di Tristan</i> in München, BSB, Cgm 51	41
I <i>Capricci d'un Re</i> digitale	48
Harmonizing and modelling a bibliographic ontology of imperial diplomas within the OntoVE Knowledge Base	51
Mappare il museo in IIIF. Una combinazione di deep zoom e VR360 per la Galleria Borghese di Roma	57
H2IOSC: Humanities and Heritage Open Science Cloud	63
Śivadharmā Database CMS. HTML and graph as a starting point for digital editions	65
L'approccio human-centered nella progettazione di un ecosistema digitale che massimizzi la diffusione del materiale diaristico preservandone la singolarità	71
Un futuro digitale per gli archivi de 'Gli Anni della Cupola'	78
<i>Ianus 2.0</i> . Corrottele filologiche, congetture digitali	85
Verso una «volontà d'archivio» digitale	92
L'applicazione del digitale alla valorizzazione della lingua come patrimonio culturale immateriale	99
Digital <i>Accrocchio</i> : a computational image searching tool for social history	104
Quali prospettive per ItaDraCor? Risorse e strumenti per la codifica di testi drammatici in lingua italiana	108
Metodologie di indagine per una nuova piattaforma: Digital Scholarly Editions Platform and Aligned Translation	111
Making Korean Perseus Digital Library: The Case Study of 'Dokdo Digital Library'	114
Constructing an Old English WordNet: The Case of <i>Guilt</i>	122
A digital diplomatic edition of the 10th-century charters of Lucca for Latin corpus linguistics	124
On why and how we should build a catalogue of software products for digital scholarly editing	130
Database Allegorico Dantesco	134
Tecnologie digitali per la realizzazione e adattamento di contenuti scolastici facilitati e semplificati	139
<i>Proteus</i> : marcatura di testi per una lettura antropologica del mito greco e romano	145
Seeing the Unseen: an approach for visualising situated statements in the Cultural Heritage domain	149
<i>Soavissime ariette</i> : impiego di repertori digitali per lo studio della tradizione delle arie d'opera	154
<i>Musisque Deoque</i> : la nuova organizzazione dei testimoni manoscritti	159
Storytelling with Linked Open Data	164
From ontology design to user-centred interfaces for music heritage	168

Un modello open data per l'interazione utente: la digital library Open Memory Apulia	173
Nel laboratorio di <i>Poeti d'Italia in lingua latina</i> : lavori in corso e prospettive future	177
Metadata e concetti, una nuova intermediazione delle fonti. Quali prospettive per la ricerca e la diffusione della storia?	181
Metodi di armonizzazione per codifiche TEI cooperative: l'edizione Bellini Digital Correspondence	185
LINEA: per un portale della letteratura italiana del Novecento	191
'Voce dei Margini': Le annotazioni di Giorgio Bassani	197
Scholarly edition of born-digital literary sources: a case study	204
Per un atlante delle biblioteche medievali: il progetto Mirabile-Atlas e una nuova prospettiva sulla circolazione della cultura manoscritta	208
Dai metadati ai dati, dai contesti ai contenuti: aumentare la descrizione archivistica	214
ITINERA: progetto, database e prospettive	220
Per l'edizione digitale delle lettere familiari di Verga: un esperimento di analisi semantica	224
Ease the collaboration making Scholarly Editions: the GreekSchools case study	230

## **2. Analisi computazionale dei testi** **233**

Zoo-KG: A knowledge graph of zoological data	234
<i>Do AIs dream of electric comics?</i> Generative AI models, digital memory, and creativity	237
ChatGPT: un Pappagallo Stocastico può essere di aiuto a un Vero Ricercatore (Umanistico)?	245
Un atlante digitale per i borghi delle aree interne	251
To what extent are lemmatisation and annotation relevant for deep learning assignments and textual motifs detection? The case-study of Peter Damian's letters (11 <sup>th</sup> century)	254

## **3. Intelligenza artificiale e modelli applicati ai beni culturali** **260**

<i>Falso movimento</i> (Moretti 2022): Memories and Translations	261
Verso la creazione della versione italiana di SPARSAR	265
Unsupervised Information Extraction from Medieval Charters	275
The challenges of sense granularity in word sense disambiguation for Latin	279
Machine Learning Algorithm for the Scansion of Old Saxon Poetry	284
Diachronic and diatopic word embeddings from British historical newspapers	289
GPT-3 vs. Delta. Applying stylometry to large language models	292
EcoLet. Ecotemario della Letteratura italiana	298
Risorse linguistiche per lo studio dei <i>Promessi Sposi</i>	301

## **4. Preservazione della memoria e del patrimonio digitale** **304**

Uso e riuso di dati per la valorizzazione del patrimonio culturale italiano online: buone pratiche per l'interoperabilità	305
"Dalla civiltà dell'archivio alla civiltà dell'algoritmo?"	309
<i>Breaking the Canvas</i> A Serious Game to challenge stereotypes in Visual Culture through Fine Arts	315
CoMontology. Towards An Ontology for the Chronology of Mediterranean archaeologies: a model for the digital memory	322
Dematerializzazione e sostenibilità: prospettive ideali e problemi concreti	327
New Media Art: indagini conservative sulle eredità immateriali dell'arte di fine millennio	332
The challenges of preservation: digitizing graffiti in the urban landscape	338
The Lions of Venice Arsenal: a neuroscientific approach to Cultural Heritage	343
L'immaginazione tra 0 e 1. Alcune considerazioni su memoria e oblio nell'era digitale	347
A methodological study on the projects developed by the Cologne Center for Humanities (CCeH)	351

Preservare il <i>Tesoro</i> . Per un nuovo sistema di interrogazione dei corpora dell'Opera del Vocabolario Italiano	358
La galassia MQDQ tra innovazione e tradizione	365
Memorizzare o ricordare? La memoria come fatto emozionale	371
GECA 3.0 – A new tool for cataloguing and enjoying cultural heritage	373
Libri, memoria e territorio: la digital library di Carlo Alberto	380
Un modello di catalogazione digitale delle grammatiche bilingue di spagnolo per italofofi	382
Un'ontologia per la classificazione semantica dei dizionari storici: l'esempio del LEI Digitale	385
How many oral archives are in your home? Piloting a new Tuscan census in the Gra.fo Reloaded project	389
Le mafie e le Digital Humanities	395
I Confini della Lunigiana: un progetto di Digital Public History	401
L'immaginario francese della Puglia: una collezione digitale del 'Petit Tour' dell'altra Italia	404
<b>5. Workshop</b>	<b>409</b>
Workshop sulle edizioni digitali: preparazione con codifica XML TEI e visualizzazione con il software EVT	410
Trascrivere, analizzare e pubblicare documenti d'archivio: Transkribus e l'intelligenza artificiale al servizio dello storico	416
Codificare (a)l'Infinito	421

## Prefazione

La dodicesima edizione del Convegno annuale di AIUCD – Associazione per l’Informatica umanistica e la cultura digitale, è intitolata *La memoria digitale: forme del testo e organizzazione della conoscenza*, con un programma che ha l’intento di interrogarsi sulle forme della testualità, intese in senso ampio e trasversale, dando spazio a riflessioni teoriche e metodologiche negli ambiti della filologia digitale, dell’analisi computazionale, dell’organizzazione e preservazione di collezioni digitali e nate digitalmente, dell’applicazione di tecniche di Intelligenza Artificiale e dei problemi della conservazione del patrimonio digitale.

La sede di Siena ha ricevuto il testimone a Lecce, durante l’XI Convegno annuale che si è tenuto presso l’Università del Salento dall’1 al 3 giugno 2022: un passaggio di consegne che ci onora e che cade nel quindicesimo anno dalla fondazione del Master in Informatica del testo ed edizione elettronica (Infotext), oggi diretto da Elisabetta Bartoli, e a venticinque anni dal pionieristico avvio delle attività digitali del Centro di Studi Comparati I Deug-Su. Quest’anno il Dipartimento di Filologia e critica delle letterature antiche e moderne dell’Università di Siena è stato inoltre riconfermato Dipartimento di Eccellenza, il progetto del quinquennio 2023-2027 include un rafforzamento delle attività di ricerca legate all’Informatica umanistica, con la prosecuzione delle numerose collaborazioni con altri dipartimenti dell’Ateneo e con enti nazionali e internazionali.

Nel 2022 Fabio Ciraci e Patrizia Miggiano avevano proposto un ricchissimo programma di incontri online, “Loading AIUCD2022”, per prepararci insieme al convegno di Lecce. Abbiamo idealmente proseguito l’iniziativa con “Aspettando AIUCD2023”, un ciclo di workshop online organizzato in collaborazione con CLARIN-IT e con il suo Knowledge Centre for Digital and Public Textual Scholarship, con il sostegno del Centro di Ricerca Interdipartimentale in Digital Humanities dell’Università del Salento, del *Journal of the Text Encoding Initiative* e di IIF. “Aspettando AIUCD2023” ha avuto come docenti dei workshop Francesco Mambrini (*Introduzione a Python per l’analisi dei dati testuali*, 4 maggio 2023), Lindà Spinazzè (*Zotero per la gestione dei dati bibliografici e la ricerca collaborativa*, 11 maggio 2023), Angelo Mario Del Grosso (*Selezione, elaborazione e presentazione di documenti XML-TEI mediante i linguaggi XPath e XSL*, 18 maggio 2023), Tiziana Mancinelli (*Introduzione a IIF – International Image Interoperability Framework*, 25 maggio 2023) e Rachele Sprugnoli (*Introduzione al Natural Language Processing*, 1 giugno 2023). I video saranno disponibili sui canali di CLARIN-IT e Università di Siena dopo la conferenza. Sono state inoltre proposte due lezioni online aperte al pubblico, in collaborazione con il Master Infotext: Klaus Kempf, *Ritorno al passato. Mostre virtuali e Iconic Turn* (13 marzo 2023), e Franco Moretti, *Semantica delle reti drammatiche* (2 maggio 2023).

La Call for Papers ha ricevuto 96 proposte da oltre duecento autrici e autori, provenienti dall'Italia e da Austria, Belgio, Brasile, Repubblica Ceca, Finlandia, Francia, Germania, Norvegia, Corea del Sud, Spagna, Svizzera, Regno Unito e Stati Uniti. La sezione "Archivi, edizioni digitali, organizzazione della conoscenza", presieduta da Marina Buzzoni, Paolo Monella e Roberto Rosselli Del Turco, ha ricevuto 47 proposte; la seconda, "Analisi computazionale dei testi", presieduta da Fabio Ciotti e Rachele Sprugnoli, 6 proposte; la terza, "Intelligenza Artificiale e modelli applicati ai beni culturali", presieduta da Monica Bianchini e Federico Boschetti, 14 proposte; la quarta, "Preservazione della memoria e del patrimonio digitale", presieduta da Nicola Barbuti e Maurizio Lana, 26 proposte; i Workshop, presieduti da chi scrive, hanno ricevuto 3 proposte, sottoposte a revisione come le quattro track della call. Sono stati accettati complessivamente 82 interventi, con una media di accettazione dell'85,4%: 47 comunicazioni orali, 32 poster, molti dei quali già disponibili su Zenodo e Figshare, e 3 workshop.

In linea con la Call for Papers e con i temi di questa edizione del convegno, il programma prevede sette relatrici e relatori invitati: il 5 giugno sarà interamente dedicato alle infrastrutture di ricerca, con una sessione presieduta da Marina Buzzoni e Cristina Marras, e interventi di Lino Leonardi e Emiliano Degl'Innocenti (*Filologia digitale e infrastrutture di ricerca, nel PNRR e oltre*), Arianna Ciula (*Infrastrutture di ricerca: dal ruolo della modellizzazione nell'ecosistema di laboratorio*), Monica Monachini e Francesca Frontini (*Scienza aperta e formazione nelle infrastrutture di ricerca europee e nazionali per le risorse linguistiche*). Il secondo giorno sarà dedicato alla filologia digitale, con l'intervento di Paola Italia, introdotta da Roberto Rosselli Del Turco, *Edizione, Annotazione, Visualizzazione. Problemi e prospettive della/per la filologia digitale*. Il 7 giugno avremo infine il contributo di Thorsten Ries, introdotto da Francesca Tomasi, con una riflessione sul trattamento degli archivi nati digitalmente: *Born-digital Literary Archives – the Present and the Future of an Emerging Field*.

Desideriamo ringraziare i Chair di track, il Comitato scientifico e i revisori che hanno generosamente contribuito, tutti coloro che hanno partecipato alla Call for Papers e i numerosi partecipanti iscritti al convegno. Un ringraziamento va ai chair di sessione che riuniranno in percorsi trasversali gli interventi: Nicola Barbuti, Paola Bellomi, Andrea Bolioli, Federico Boschetti, Flavia Bruni, Paola Castellucci, Fabio Ciraci, Angelo Mario Del Grosso, Giulia Giovani, Maurizio Lana, Paolo Monella, Daria Spampinato, Rachele Sprugnoli. Ringraziamo il Comitato Scientifico e il Comitato Organizzatore per il loro essenziale contributo, il Direttivo di AIUCD per la costante collaborazione, CLARIN-IT per il generoso supporto. Un ringraziamento speciale va alla Segreteria Amministrativa del DFCLAM, e in particolare a Francesca Pietrini; a Gabriele Lo Piccolo, Alessia Valenti, e a tutte e tutti i giovani che hanno collaborato e collaboreranno nei giorni della conferenza: Giulia Bassi, Silvia Cappa, Chiara Cauzzi, Martina Corti, Elena Crocicchia, Anna Guadagnoli, Giada Giannetti, Bogdan Groza, Alessia Luvisotto, Paola Mocella, Pietro Orlandi, Martina Paccara, Elisa Petri, Maria Grazia Schiaroli.

Questo convegno sarà particolarmente difficile, il nostro pensiero è rivolto a Dino Buzzetti: la sua improvvisa scomparsa, il 23 aprile 2023, ha profondamente toccato tutta la nostra Associazione. Ma sarà anche un momento che riunirà le socie e i soci per confermare l'impegno di tutti noi nell'operare ricordandoci il suo insegnamento, il suo rigore scientifico e il suo spirito di servizio uniti all'estrema generosità e gentilezza. Con il Direttivo di AIUCD desideriamo che questi giorni siano prima di tutto un'occasione per ricordare con profonda riconoscenza il nostro Presidente onorario, tra i fondatori dell'Associazione nel 2011: il 5 giugno, durante la cerimonia di inaugurazione, Francesca Tomasi ci consegnerà un ricordo che abbiamo anticipato in questo volume; il 6 giugno intervengono Daniela



Calanca e Claudia Buzzetti, insieme ad Anna Maria Tammaro; infine, in chiusura di convegno, avremo un contributo di Fabio Ciotti, a cui idealmente uniamo i numerosi messaggi di gratitudine pubblicati in queste settimane sul blog dell'Associazione.

Emmanuela Carbé

Francesco Stella

## In ricordo di Dino

Francesca Tomasi



Foro rivista Ecdotica, 2004

Come ho avuto modo di ricordare durante la celebrazione in Certosa con cui abbiamo salutato Dino, era il 1993 quando l'ho conosciuto. Conosciuto come professore del corso di Informatica Documentale della Facoltà di Conservazione dei Beni culturali di Ravenna. E come ricordavo in quella occasione, ovvero lo scorso 28 Aprile, sono state le sue lezioni a farmi capire, appena ventenne, cosa avrei voluto fare nella vita. Quale sarebbe stato il campo di ricerca a cui avrei voluto dedicare i miei sforzi. E non posso non ricordare il mio primo lavoro, avviato con Dino nel 1997 per la tesi di laurea, sulla creazione di un'edizione elettronica in SGML/TEI, ovvero la rappresentazione della struttura narratologica delle *Novelle Porrettane* di Sabadino degli Arienti visualizzata attraverso una primissima versione sperimentale del tool Anastasia di Peter Robinson (versione ufficiale del 2000) e mostrata alla commissione di tesi con un enorme calcolatore che mi ero portata direttamente da casa. Incontrare un docente capace di far prendere coscienza così presto del proprio futuro è una fortuna che non capita a tanti. Auguro, in particolare ai più giovani, che per loro questo incontro magico possa avvenire (o auspico sia avvenuto grazie ai tanti validi colleghi che vedo qui presenti oggi).

Dino è stato un intellettuale raffinato, con un incredibile spirito critico, e soprattutto un visionario, precursore di tante ricerche, avviate in modo organico molti anni dopo rispetto ai suoi primissimi studi.

Una mente in grado di riflessioni teoriche di enorme spessore. E capace di innestare nell'interlocutore un modo nuovo di osservare la realtà, di porsi nuove domande di ricerca. Abile nello stimolare un'apertura di pensiero, con una preziosissima, innata, abilità, davvero fuori dal comune. Un ispiratore, direi.

I suoi primi scritti degli anni '90 sono stati l'avvio di un approccio nuovo alle discipline umanistiche e alla ricerca mediata dal calcolatore, alle pratiche computazionali come strategia di modellazione del testo. E penso in particolare al contributo con Andrea Tabarroni, *Informatica e critica del testo: Il caso di una tradizione 'fluida'*<sup>1</sup> del 1991 e a D. Buzzetti, P. Pari, A. Tabarroni, *Libri e maestri a Bologna nel XIV secolo: Un'edizione come database*.<sup>2</sup> Ovvero i primi esperimenti condotti con Manfred Thaller sul sw Kleio. E senza dimenticare, di qualche anno dopo, il fondamentale: *Rappresentazione digitale e modello del testo*, in *Il ruolo del modello nella scienza e nel sapere*<sup>3</sup> negli atti dei Convegni Lincei del 1998.

Rileggendo oggi i suoi scritti, sono incappata nel contributo *il sogno e la precisione*, con cui Dino ricorda Gigliozzi, nel volume curato da Mordenti in suo onore,<sup>4</sup> e così dice: «L'umanità e l'ironia di Giuseppe, sempre accompagnate da quello sguardo sereno, rendono difficile immaginarne l'intransigenza. Ma c'è intransigenza nel rispetto. E a Giuseppe l'intransigenza veniva dal rispetto per la qualità intellettuale e scientifica del suo lavoro. *Non perché fosse il suo, ma perché quel suo lavoro implicava un atto di umiltà*».

E mentre leggo queste parole scritte per Gigliozzi mi sembrano appropriate anche per Dino. Intellettuale fuori dalle logiche accademiche e dai giochi di palazzo. Sempre più orientato alla ricerca come aspirazione alla conoscenza che non alla carriera. Sempre attento anche ai più giovani, che ascoltava con interesse durante i nostri convegni e con i quali dialogava alla pari, senza mai far pesare il suo ruolo o le sue competenze.

La sua ricerca dal 2000 in poi è stata fondamentale per la costituzione delle basi teoriche della *Humanities Computing* come la conosciamo oggi. Anche a livello internazionale. Con contributi scientifici fondativi come: *Codifica del testo e intelligenza artificiale*<sup>5</sup> (che, notate, è del 2003); *Diacritical Ambiguity and Markup*<sup>6</sup>; i lavori con McGann, come *Critical Editing in a Digital Horizon*, nel volume *Electronic Textual Editing*<sup>7</sup>; o ancora *Digital Editions and Text Processing* in M. Deegan and K. Sutherland (eds.), *Text Editing, Print, and the Digital World*<sup>8</sup>; fino agli scritti più recenti<sup>9</sup>.

A lui dobbiamo infatti anche la fondazione della nostra Associazione AIUCD nel 2011 (di cui è stato Presidente ed è, sempre sarà, presidente onorario) e tanto dobbiamo alla sua lungimiranza rispetto alle strategie più efficaci per dare giusto valore e riconoscimento al nostro campo di studi.

---

<sup>1</sup> Uscito su «Schede umanistiche», n.s., 5:2 (1991), pp. 185-193.

<sup>2</sup> Sempre su «Schede umanistiche», n.s., 6:2 (1992), pp. 163-169.

<sup>3</sup> Atti del Convegno (Roma, Accademia Nazionale dei Lincei, 27-28 ottobre 1998), Roma, Accademia Nazionale dei Lincei, 1999, (Contributi del Centro Linceo Interdisciplinare «Beniamino Segre», N. 100), pp. 127-161.

<sup>4</sup> D. Buzzetti, *Il sogno e la precisione*, in R. Mordenti (a cura di), *Giuseppe Gigliozzi: La fondazione dell'informatica applicata al testo letterario* («Testo & senso», 4/5, 2001-2002), Roma, EUROMA-Editrice Universitaria di Roma «La Goliardica», 2002, pp. 24-28.

<sup>5</sup> Sempre su «Schede Umanistiche» n.s., 17:1 (2003), pp. 171-197.

<sup>6</sup> In D. Buzzetti, G. Pancaldi, and H. Short (eds.), *Augmenting Comprehension: Digital Tools and the History of Ideas*, London-Oxford, Office for Humanities Communication, 2004.

<sup>7</sup> Edited by L. Burnard, K. O'Brien O'Keefe, and J. Unsworth, New York, The Modern Language Association of America, 2006, pp. 51-71.

<sup>8</sup> Aldershot, Ashgate, 2009, pp. 45-62.

<sup>9</sup> D. Buzzetti, *Dell'ineluttabile "paradossalità" del testo: La conferma digitale*, in *Letteratura e altre rivoluzioni: Per Raul Mordenti*, a cura di Domenico Fiormonte e Paolo Sordi, Roma, Bordeaux, 2020, pp. 198-229; D. Buzzetti, A. De Ninno, D. Fiormonte, *Informatica Umanistica e Cultura Digitale: La sfida epistemologica*, in *AIUCD 2021 - DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale*, a cura di F. Boschetti, A. M. Del Grosso ed E. Salvatori, Pisa, AIUCD, 2021, pp. 319-326.

Ma Dino era anche una persona che sapeva divertirsi, con cui ci stava volentieri in compagnia, che condivideva tanto il momento della ricerca, quanto quello dello svago. E, aggiungerei, è stato una presenza importante anche nel momento del bisogno personale. Più che un maestro, direi un confidente e davvero un amico.

Chiudo dicendo allora che credo sia urgente pensare, come già abbiamo iniziato a fare con il collega Ciotti, alla raccolta dei suoi scritti, operazione che già doveva essere un regalo per i suoi 80 anni e che non è andata in porto (*mea culpa*). È un'azione necessaria, perché chiunque voglia fare ricerca in questo settore della IU non può non aver letto le cose che ha scritto il prof. Buzzetti.

Permettetemi un'ultima nota personale. Ricordatevi di far sapere alle persone che stimate quanto le apprezzate, quanto riconoscete il loro valore, quanto è preziosa per voi la loro mente brillante. Per non pentirvi di non averlo fatto abbastanza quando avreste potuto...

Ciao Dino, sempre con noi.

Francesca

## **1. Archivi, edizioni digitali, organizzazione della conoscenza**

# Creare e interrogare i modelli: un duplice valore conoscitivo. Il caso dei pareri di lettura einaudiani di Natalia Ginzburg

Laura Antonietti<sup>1</sup>

<sup>1</sup> Université Grenoble Alpes, France – laura.antonietti@univ-grenoble-alpes.fr

## ABSTRACT

L'articolo presenta i risultati della modellizzazione e dell'analisi del processo di lettura all'interno della casa editrice italiana Einaudi nel secondo dopoguerra, con un'attenzione particolare a Natalia Ginzburg (1916-†1991). Più specificamente, l'intento dell'intervento è mostrare come gli strumenti delle *Digital Humanities* abbiano contribuito a realizzare gli obiettivi scientifici del lavoro di ricerca. Da un lato, le *Digital Humanities* hanno permesso di modellizzare, rappresentare e interrogare il corpus dei documenti in modo pertinente, efficace ed efficiente; dall'altro, hanno dato un contributo fondamentale e indispensabile dal punto di vista metodologico, euristico e interpretativo.

## PAROLE CHIAVE

Modellizzazione; UML; database; XML TEI; archivi editoriali; pareri di lettura

## 1. INTRODUZIONE

Nata nel 1933 con una fisionomia militante e antifascista, Einaudi diventa una delle principali case editrici di cultura nel secondo dopoguerra. Tra i protagonisti di questo successo troviamo la scrittrice e redattrice Natalia Ginzburg, che ha contribuito a costruire l'identità della casa editrice e il suo catalogo, in particolar modo in qualità lettrice di opere narrative contemporanee (Antonietti 2022 e 2021, Iannuzzi 2012, Saita 2009). Modellizzare il processo di lettura einaudiano ci permette di porre in evidenza la centralità di Ginzburg, una vera e propria «lettrice formidabile» (Cesari 1991), all'interno dei meccanismi decisionali e di metterne in luce le ragioni. Ci proponiamo quindi in primo luogo di presentare il corpus dei documenti, in seguito di illustrare i passaggi della costruzione dei modelli che ci permettono di rappresentare e interrogare i dati e infine di riflettere sul contributo metodologico ed euristico dell'attività di modellizzazione.

## 2. IL CORPUS DELLE FONTI

Il corpus dei documenti è costituito dai pareri di lettura einaudiani relativi alle opere narrative coeve, redatti tra gli anni '40 e gli anni '60: essi rappresentano una fase fondamentale del processo decisionale che porta alla pubblicazione di un'opera e quindi, in ultima analisi, alla costruzione del catalogo di una casa editrice. Tra gli estensori di questi documenti figurano alcuni tra i più significativi scrittori e intellettuali del Novecento (Cesare Pavese, Elio Vittorini, Italo Calvino, Natalia Ginzburg, solo per citarne alcuni), che giudicano le prove narrative di altri autori, confrontandosi allo stesso tempo con la propria idea di letteratura. I pareri di lettura non sono quindi semplici documenti di lavoro, ma esempi di meta-letteratura che costituiscono un vero e proprio genere critico (Bricchi 2010), che rivelano i modelli letterari stabiliti da Einaudi e dai suoi collaboratori e che devono essere interpretati secondo un preciso orizzonte storico e culturale (Cadioli 1995). Ci permettono inoltre di approfondire la conoscenza della figura intellettuale dei lettori che li hanno redatti, documentando il funzionamento e il carattere eccezionale del laboratorio intellettuale che l'Einaudi nel dopoguerra ha costituito.

Nonostante la loro ricchezza, i documenti relativi al processo di lettura rimangono poco studiati e in gran parte inediti. Si tratta di lettere, schede, verbali di riunioni e comunicazioni editoriali di vario tipo, che costituiscono una documentazione complessa, caratterizzata dall'eterogeneità e dalla multi-autorialità, nonché dalla quantità di attori coinvolti, spesso con ruoli diversi. Il corpus raccolto testimonia quindi un metodo di lavoro orizzontale e collegiale, tanto da far parlare la storiografia e la critica di «casa-laboratorio» o, ancora, di «cervello collettivo» (Ferretti 2004).

## 3. LE FASI DELLA MODELLIZZAZIONE

Data questa complessità, la sfida principale è quella di descrivere e rappresentare, e quindi modellizzare, il processo di lettura di Einaudi attraverso i documenti prodotti all'interno di esso. Da un punto di vista metodologico, la modellizzazione ha coinvolto tre aspetti differenti, per ognuno uno dei quali è stato scelto un apposito strumento: la delimitazione del processo

di lettura editoriale nella sua collegialità e pluralità di attori coinvolti (UML), la descrizione delle caratteristiche fondamentali di tale processo e dei documenti prodotti (Database PostgreSQL) e la rappresentazione del contenuto e la struttura di questi ultimi (XML TEI).

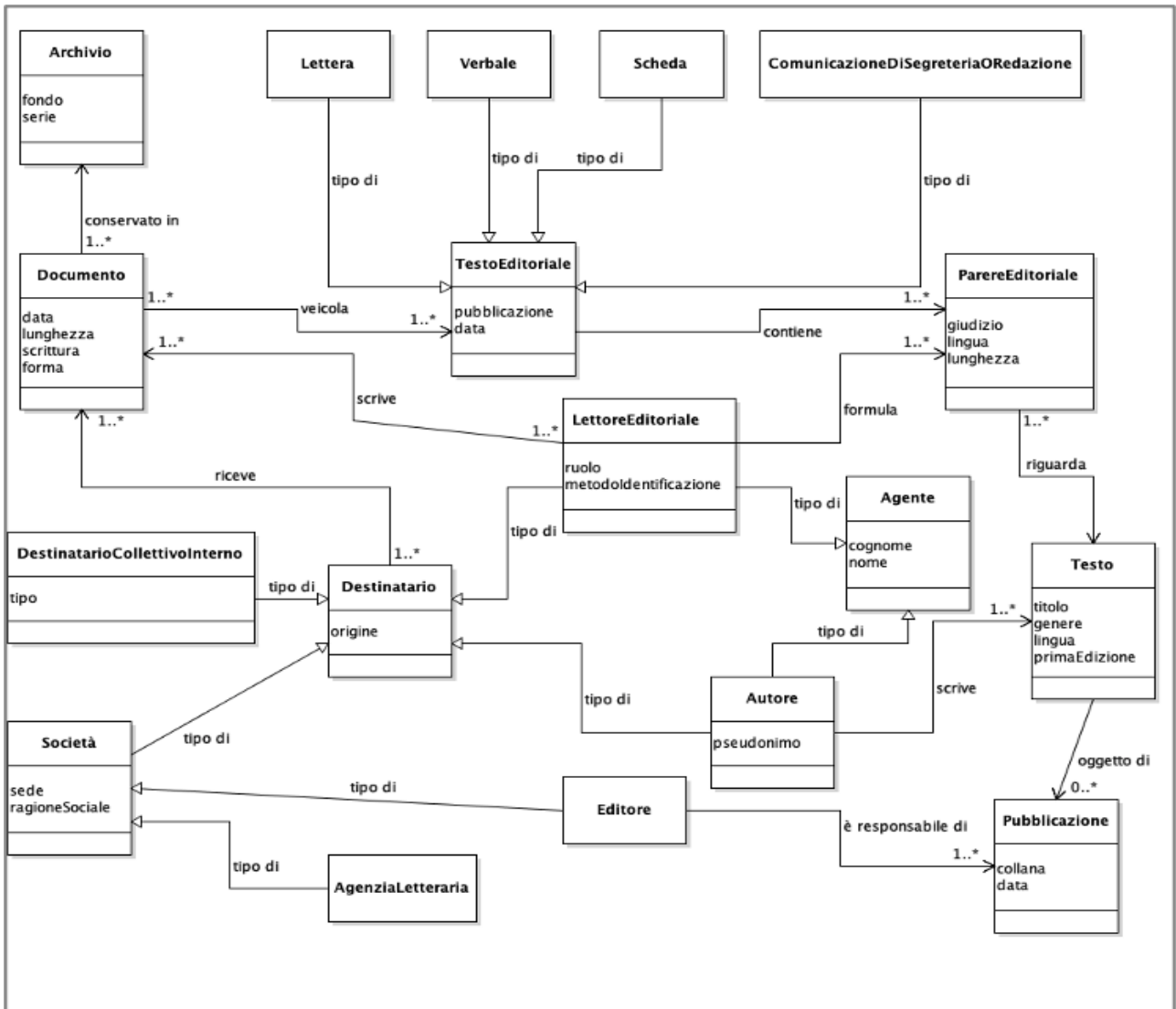


Figura 1. Il diagramma delle classi UML.

Per creare il modello concettuale atto a descrivere il processo di lettura editoriale einaudiano (fig. 1), è stato scelto di utilizzare il linguaggio UML (Unified Modeling Language). Riprendendo in larga misura le idee fondamentali dell'*entity-relationship model*, UML permette di realizzare diagrammi che identificano i concetti chiave di un determinato dominio e le relazioni esistenti tra essi. Essendo al contempo uno strumento di *visual modeling*, UML consente la visualizzazione grafica dell'architettura e della struttura del modello realizzato nel dettaglio, utilizzando un insieme ben definito di elementi grafici. Tra le molte opzioni di diagramma offerte, è stato scelto di utilizzare un *diagramma delle classi*, che da un lato ha permesso di rappresentare il processo di lettura mantenendo il focus sui documenti da esso prodotti, sugli agenti coinvolti e sulle relazioni esistenti tra essi, dall'altro ha costituito le basi per il lavoro di costruzione del database delle fonti. Il linguaggio UML si è rivelato il più efficace per produrre lo schema adatto a descrivere nel modo più completo e coerente possibile il processo di lettura einaudiano. L'atto di dare un nome alle entità presenti, formalizzandole, nonché alle loro relazioni, ha avuto infatti il ruolo decisivo di fare luce e chiarezza sulla complessità di tale processo: le regole sintattiche precise e definite di un linguaggio semi-formale come UML hanno dato una forma di rigore e strutturazione al pensiero; la libertà semantica ha, al contrario, permesso di descrivere in modo idoneo e specifico il dominio d'interesse.

Il diagramma UML è servito come base per la costruzione di un database che consente l'inventario delle fonti e l'interrogazione dei dati e delle relazioni esistenti tra di essi (fig. 2). Si tratta di un'applicazione creata utilizzando il framework Django, installato su un server Heroku, che fornisce un database relazionale open source realizzato con



PostgreSQL. La piattaforma genera automaticamente l'interfaccia web per l'inserimento dei dati dalle tabelle del database, progettate a partire dalle classi UML. Si tratta di uno strumento flessibile e ampliabile, sia da un punto di vista tecnico, in quanto l'uso di tecnologie e linguaggi gratuiti e open source, così come di standard comuni e condivisi, garantiscono l'interoperabilità e l'esportabilità dei dati, sia da un punto di vista concettuale, per la descrizione e l'analisi di processi decisionali e di tipologie documentarie differenti, al di là del caso di studio einaudiano. Dal database è possibile accedere anche alla trascrizione e alla codifica dei documenti, che seguono le linee guida TEI (fig. 3).

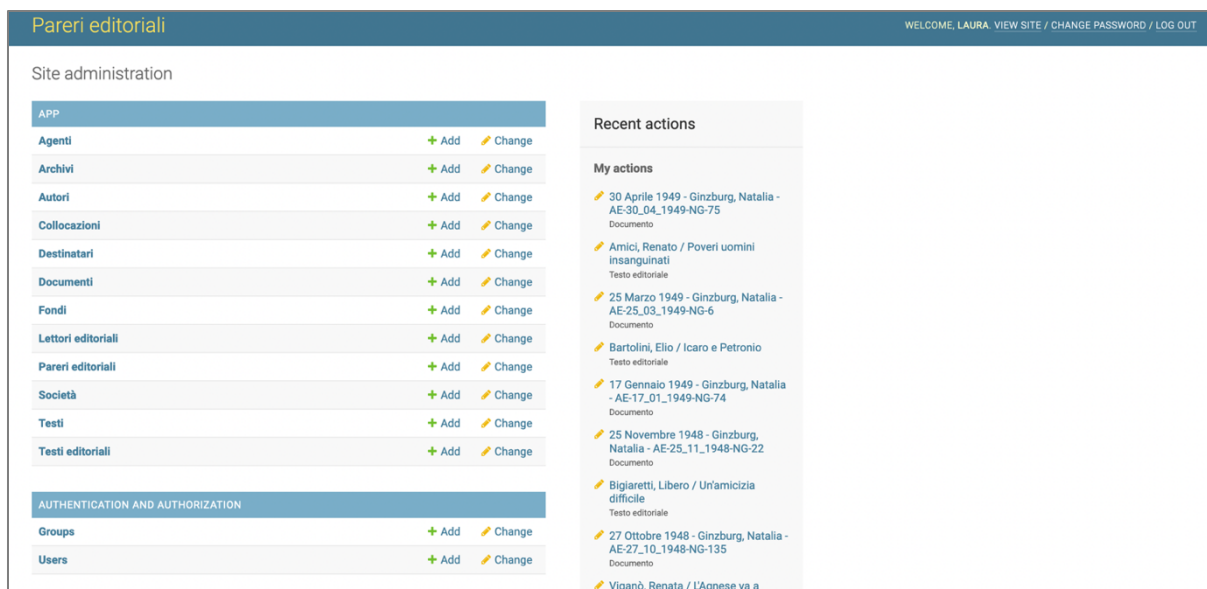


Figura 2. Database - homepage.

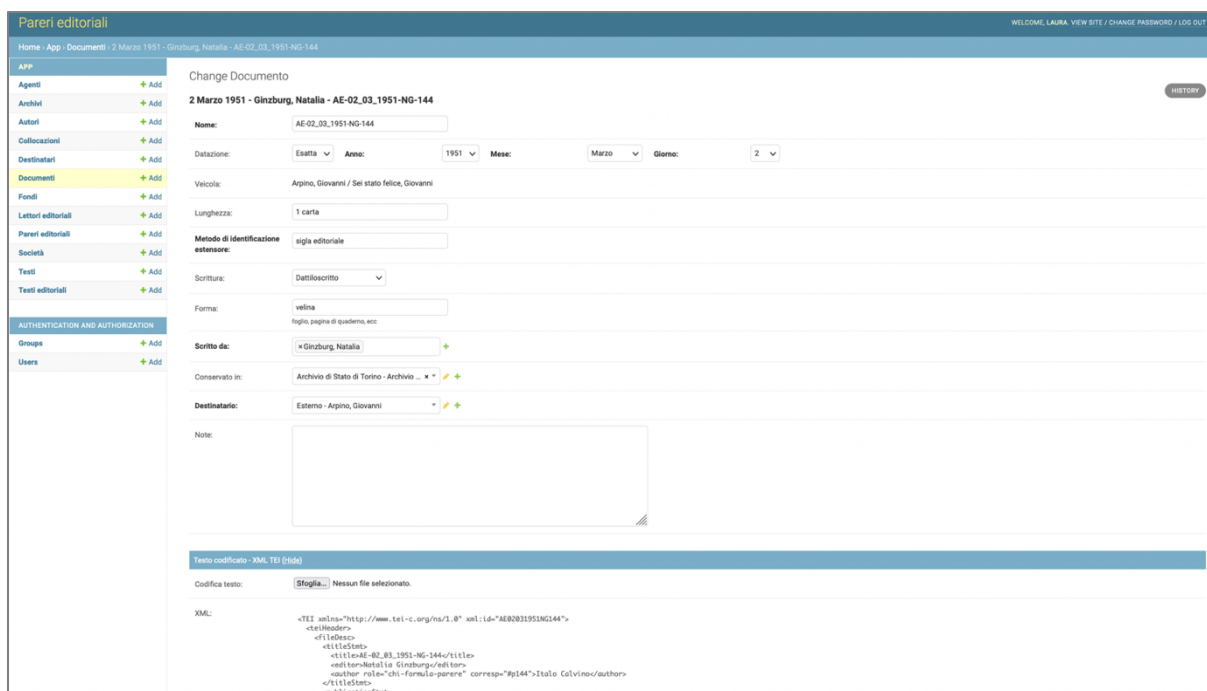


Figura 3. Database - codifica del testo.

In parallelo all'inventario delle fonti e alla raccolta dei metadati attraverso il database, è stata infatti condotta un'operazione di trascrizione e codifica del corpus. È stato elaborato un modello che mette in evidenza le diverse tipologie testuali, la loro struttura e le relazioni esistenti tra essi. Non si è trattato di un'operazione di modellizzazione evidente, in quanto la TEI non possiede un modulo specifico per la codifica dei pareri di lettura: il modello creato rimanda quindi a diversi moduli esistenti. I marcatori che si è deciso di utilizzare e le relative porzioni testuali che sono state sottoposte a codifica sono i seguenti:

- **<persName>**: nomi di persona. Attraverso l'attributo @role, è stato possibile identificare la funzione dell'agente in questione (autore, editore, collaboratore editoriale).
- **<placeName>**: nomi di luogo, in particolare di città (utili da identificare nell'ottica della ricostruzione del dialogo tra le diverse sedi einaudiane).
- **<orgName>**: nomi di organizzazione. Attraverso l'attributo @type, è possibile distinguere le varie tipologie (casa editrice, agenzia letteraria, ecc.).
- **<opener>** e **<closer>**: struttura di una lettera, apertura e chiusura.
- **<div>**: blocchi testuali, che permettono di isolare i diversi pareri e l'eventuale presenza di diversi testi editoriali (attraverso l'attributo @type).
- **<metamark>**: sigle editoriali, che consentono di identificare l'estensore del documento in mancanza di una firma.
- **<del>** e **<add>**: eventuali cancellazioni o aggiunte significative. I due elementi vengono inseriti all'interno del marcatore <subst> se, in unico intervento di riscrittura, una porzione di testo viene cancellata e un'altra viene aggiunta.
- **<sic>** e **<corr>**: errori o sviste nella fonte e correzione. Entrambi gli elementi sono compresi all'interno dell'elemento <choice>.
- **<supplied>**: integrazioni del testo mancante nella fonte.
- **<unclear>**: letture incerte, dovute a difficoltà di decifrazione della grafia. Qualora sia risultato impossibile persino ipotizzare quanto sia stato scritto, si è utilizzato l'elemento <gap>, con "illeg" come valore dell'attributo @reason.

I testi sono stati codificati manualmente, mentre i metadati relativi alla descrizione contenuti nel <teiHeader> sono stati generati automaticamente da uno script a partire dalle tabelle del database, secondo il seguente modello:

- **<filedesc>**: contiene una descrizione bibliografica completa del documento digitale. A sua volta comprende:
  - **<titleStmt>**: titolo del documento e responsabilità del suo contenuto intellettuale.
  - **<publicationStmt>**: informazioni relative alla pubblicazione e distribuzione di un testo, elettronico o meno.
- **<sourceDesc>**: dettagli relativi alla o alle fonti da cui è derivato il documento elettronico. Contiene <msDesc>, che comprende la descrizione di diversi aspetti della fonte manoscritta o dattiloscritta:
  - **<msIdentifier>**: la sua collocazione archivistica.
  - **<msContent>**: il suo contenuto, che include un breve riassunto e l'eventuale pubblicazione del testo oggetto di lettura o, ancora, il riferimento bibliografico di un'eventuale pubblicazione del testo editoriale.
  - **<physDesc>**: descrizione del documento. Contiene a sua volta:
    - **<objectDesc>**: descrizione dell'oggetto fisico, ossia del supporto e della sua estensione.
    - **<handDesc>**: descrizione del tipo di scrittura.
    - **<history>**: data in cui il testo editoriale è stato redatto.
- **<profileDesc>**: descrizione dettagliata degli aspetti non bibliografici relativi al documento. In particolare, i giudizi espressi nei testi editoriali e la lingua utilizzata.

Questa scelta ha permesso, da un lato, di evitare fenomeni di ridondanza ed occasioni di errore e, dall'altro, di accelerare e ottimizzare il processo di trascrizione e codifica. Il risultato è un catalogo riccamente annotato, in cui i metadati descrivono in dettaglio i documenti che rappresentano le relazioni tra Ginzburg e i suoi interlocutori; la consistenza dei materiali d'archivio e la loro descrizione fisica; l'elenco dei destinatari e dei mittenti delle lettere, la loro datazione; la localizzazione dei documenti (e delle copie dei documenti), fornendo un'informazione ampia e capillare sulla localizzazione dei materiali in sedi archivistiche diverse e rendendo possibili ricerche incrociate su dati presi da cataloghi differenti, consultabili spesso solo in versione cartacea o non informatizzati. Questo progetto risponde all'esigenza diffusa tra i ricercatori di poter individuare (e, nel caso della digitalizzazione, consultare) le fonti primarie in modo più semplice e rapido, grazie a cataloghi e inventari online, banche dati e strumenti di mappatura di archivi pubblici e privati.

#### 4. I RISULTATI

Lo strumento creato, che definiamo un «knowledge site» secondo la terminologia proposta da Shillingsburg (2009), merita alcune riflessioni di natura euristica. Il contributo epistemologico dato dall'attività di modellizzazione (Flanders e Jannidis

2018, Ciula e Eide 2014, Cummings 2008) è stato essenziale all'interno della ricerca condotta: questa affermazione è vera sia per la prima fase del lavoro, che potremmo definire esplorativa e conoscitiva, sia per la seconda fase, di analisi qualitativa e di interpretazione delle fonti; sia per la creazione dei modelli (il diagramma UML, le tabelle PostgreSQL e lo schema di codifica TEI XML) sia per il loro utilizzo. I linguaggi e i metodi delle *Digital Humanities* hanno infatti apportato un doppio valore di conoscenza: da un lato, hanno garantito il rigore e la strutturazione del pensiero necessari per l'organizzazione e la rappresentazione del processo di lettura e dei documenti prodotti all'interno di esso; dall'altro il risultato del nostro lavoro trova nella dinamicità e nell'interazione i suoi presupposti e le sue peculiarità: esso non potrebbe essere stampato e letto su supporto cartaceo, senza perdita di contenuto e di funzionalità. Il sistema di filtri (fig. 4) adottato ne è un esempio perfetto.

The screenshot shows a web application interface for 'Pareri editoriali'. The top navigation bar includes 'Home - App - Pareri editoriali' and user information 'WELCOME, LAURA, VIEW SITE / CHANGE PASSWORD / LOG OUT'. A sidebar on the left lists various categories with '+ Add' buttons: Agenti, Archivi, Autori, Collaboratori, Destinatari, Documenti, Fondi, Lettere editoriali, Pareri editoriali (highlighted), Società, Testi, and Testi editoriali. Below this is an 'AUTHENTICATION AND AUTHORIZATION' section with 'Groups' and 'Users' options. The main content area is titled 'Select Parere editoriale to change' and features a search bar and a table of records. The table has columns: NOME, LETTERE EDITORIALI, TESTO, AUTORI, EDIZIONI PRECEDENTI, EDIZIONE, TIPO, DESTINATARIO, GIUDIZIO, DATA, DOCUMENTO, and TESTO EDITORIALE. The filter panel on the right includes sections for 'By Lettere editoriale', 'By Opera', 'By Autore', 'By Editore', 'By Parere pubblicato', and 'By Lingua del testo proposto'.

Figura 4. Database - sistema dei filtri.

Riportiamo qui alcuni esempi di interrogazione più o meno complessi allo scopo di mettere in luce le potenzialità dello strumento creato:

- Tutti i pareri su un testo X;
- Tutti i pareri su un autore Y;
- Tutti i pareri di un lettore Z;
- Tutti i pareri di un lettore Z, su un autore Y;
- Tutti i pareri di un lettore Z, su un testo X;
- Tutti i pareri di un lettore Z, su testi poi pubblicati da un editore J;
- Tutti i pareri che esprimono un giudizio K, su un testo X;
- Tutti i pareri che esprimono un giudizio K, di un lettore Z, su testi poi pubblicati da un editore J;
- Tutti i pareri di un lettore Z, su un testo X (o su un autore Y), che esprimono un giudizio K;
- Tutti i pareri che esprimono un giudizio K, di un lettore Z, che sono stati (o meno) oggetto di pubblicazione;
- Tutti i pareri che esprimono un giudizio K, di un lettore Z, su un autore Y, che sono stati (o meno) oggetto di pubblicazione;
- Tutti i pareri che esprimono un giudizio K, di un lettore Z, che sono stati oggetto (o meno) di pubblicazione e che riguardano un testo scritto in una lingua H.

Combinare i risultati dell'applicazione dei filtri tra di loro, ottenendo informazioni più complesse, offre un importante strumento di ricerca e di interrogazione dei metadati associati al corpus. Una volta filtrati i risultati è possibile, grazie alla visione sinottica offerta dalle diverse colonne (fig. 4), integrare ulteriori informazioni sui pareri di lettura e i testi oggetto di valutazione: la data del parere, la data dell'eventuale pubblicazione del testo proposto, l'eventuale presenza di edizioni precedenti (come nel caso dei testi stranieri), la tipologia del testo editoriale e il suo destinatario.

Grazie allo strumento creato, è possibile ricostruire e rappresentare il funzionamento del laboratorio Einaudi: il suo valore in termini di conoscenza non è solo nel risultato (nella sua fruizione e interrogazione), ma anche nell'intero processo attraverso cui vi si è giunti. La modellizzazione apporta il suo massimo valore conoscitivo in quanto processo in divenire continuo. Essa presenta infatti un doppio movimento: dal concreto all'astratto (dal corpus alla sua rappresentazione concettuale – nel caso di studio qui presentato, dai pareri di lettura al diagramma delle classi e alla cernita degli elementi testuali da sottoporre ad analisi) e dall'astratto al concreto (dalla rappresentazione concettuale all'istanza modellizzata – dal diagramma delle classi e dalla griglia analitica e descrittiva elaborata alla base di dati e ai file XML TEI). Non si tratta di un percorso unidirezionale e lineare, ma di un processo dinamico e iterativo (Ciula ed Eide 2014), che evolve a seconda delle esigenze del progetto e che in alcuni casi permette di far emergere nuove domande di ricerca.

Nel caso dello studio qui presentato, queste ridefinizioni hanno riguardato, in alcuni occasioni, solo determinati aspetti della modellizzazione, mentre in altri l'hanno coinvolta a livello concettuale, logico e fisico. Da un punto di vista squisitamente implementativo, senza che le modifiche abbiano avuto conseguenze sui modelli concettuali, possiamo portare l'esempio dell'introduzione dell'insieme di filtri e della vista sinottica delle informazioni fondamentali attraverso un sistema di colonne nel database: gli elementi da visualizzare sono stati selezionati progressivamente a seconda delle ricerche che ci si è accorti di condurre più frequentemente nella base di dati. Un esempio ancora più specifico può essere la presenza della colonna "edizioni precedenti", che si è ritenuto opportuno inserire quando si è deciso di analizzare il processo di valutazione dei romanzi della narrativa francese contemporanea, data l'importanza che questa ha avuto nell'Einaudi del dopoguerra e nelle letture editoriali di Ginzburg: questo ha portato a renderci conto della rilevanza dei rapporti di Einaudi con i singoli editori d'oltralpe, attraverso una serie di risultati quantitativi che hanno poi permesso di procedere all'analisi qualitativa e alle riflessioni sul corpus.

Altre riflessioni hanno condotto a modifiche più sostanziali dei modelli dei dati e dei metadati, tanto a livello concettuale quanto a livello logico-fisico. Solo esplorando gli archivi si è potuto, ad esempio, comprendere che esistevano più copie dello stesso testo editoriale, ma con caratteristiche che potevano differire da un testimone all'altro. Questo si è tradotto nella modifica del diagramma delle classi e, quindi, della base di dati, mettendo a punto, anche a livello informatico ed implementativo, la differenza concettuale che intercorre tra dimensione documentaria e dimensione testuale; anche il modello di codifica è stato rivisto in modo che tenesse conto delle eventuali aggiunte, manoscritte o dattiloscritte.

Procedendo ulteriormente nella raccolta e nello studio delle fonti, è stato possibile rendersi conto della frequente difficoltà (se non impossibilità) di prendere visione degli originali delle lettere, conservati negli archivi dei riceventi disseminati sul territorio italiano, qualora accessibili o esistenti: non è il caso delle diverse decine di autori, le cui proposte di pubblicazione sono state considerate in questa sede, che non hanno avuto alcun futuro editoriale. Le veline, che era prassi conservare e archiviare in Einaudi, custodite (non sistematicamente, ma sovente) nei fascicoli dei mittenti, hanno rappresentato, nel contesto delle nostre ricerche, una fonte più facile da consultare e, di conseguenza, da trascrivere e codificare. Il modello di codifica, d'altro canto, ha subito ulteriori riadattamenti, partendo dalla constatazione, possibile solo in una fase più avanzata della ricerca e dell'analisi dei documenti, che le discrepanze presenti tra veline e originali (qualora disponibili) non sono così frequenti e soprattutto non significative. Questo si spiega, da una parte, considerando le peculiarità del genere del parere di lettura: l'immediatezza e la rapidità proprie di un tipo di scrittura professionale, la quale non richiede, in linea di principio, una particolare elaborazione (e rielaborazione).

D'altra parte, bisogna tenere conto di una serie di caratteristiche proprie della scrittura editoriale di Natalia Ginzburg. In primo luogo, non sapeva scrivere a macchina e delegava tale mansione alle segretarie della casa editrice, come lei stessa dichiara a più riprese: in mancanza di un originale autografo, ogni interesse di indagine del tipo sopra descritto, decade. Si è dunque deciso, in questi casi, di non codificare le eventuali cancellature e aggiunte dattiloscritte, così come di correggere in modo tacito gli errori di battitura, le sviste ortografiche e la grafia inesatta dei nomi propri. La presenza di lettere manoscritte si intensifica a partire dal 1952, quando Natalia Ginzburg lascia la sede torinese di Einaudi per trasferirsi a Roma. Ginzburg era solita scrivere a mano le proprie lettere, di getto, senza l'utilizzo di brutte copie: analizzando i documenti del corpus, è possibile accorgersi che aggiunte e cancellature sono per lo più ascrivibili alle riformulazioni proprie di una scrittura colloquiale e informale e dello stile epistolare di Ginzburg: una codifica più granulare e dettagliata è stata riservata a queste situazioni, come nell'eventualità di aggiunte e modifiche testuali effettuate in un secondo atto scrittoriale o ad opera di un altro lettore. La codifica TEI si mostra, in tal senso, uno strumento flessibile e adattabile alle specificità dei documenti presi in considerazione.

Lungi, quindi, dal rappresentare un mero esercizio accademico puramente speculativo, l'attività di modellizzazione nelle sue diverse fasi e l'interrogazione dei modelli concepiti, intese come uno strumento riflessivo e conoscitivo, hanno permesso di porre le basi essenziali per l'interpretazione critica del corpus e di giungere ai risultati critici attesi: mettere in luce un genere critico poco conosciuto (i pareri di lettura), facilitare la comprensione di documenti fortemente intertestuali e caratterizzati dalla multi-autorialità, evidenziare la dimensione collaborativa e collegiale della redazione di Einaudi, fare

luce sull'attività editoriale di Natalia Ginzburg in Einaudi, riconoscendone la centralità nei processi decisionali, nonché nella costruzione del catalogo e dell'immagine della casa editrice.

## BIBLIOGRAFIA

- [1] Antonietti, Laura. Lire des romans chez Einaudi : le cas de Natalia Ginzburg (1944-1964). Les rapports de lecture au prisme des Humanités Numériques, Tesi di dottorato discussa nell'a.a. 2021-2022, Université Grenoble Alpes – Università degli Studi di Milano.
- [2] Antonietti, Laura. «“Una lettrice formidabile”: Natalia Ginzburg e la casa editrice Einaudi». *Cahiers d'études italiennes*, 32 (2021). <https://journals.openedition.org/cei/8590>.
- [3] Booch Grady, Jacobson Ivar e Rumbaugh James. *The unified software development process: the complete guide to the unified process from the original designers*, Reading, Addison-Wesley 1998.
- [4] Bricchi, Mariarosa. «La scheda di lettura come micro-genere letterario». In S. Arduini e I. Carmignani (a cura di), *Giornate della traduzione letteraria. Nuovi contributi*. Roma: Iacobelli, 2010, pp. 54-66.
- [5] Cadioli, Alberto. *Letterati editori*, Milano: Il Saggiatore, 1995.
- [6] Cesari, Severino. *Colloquio con Giulio Einaudi*, Roma-Napoli: Theoria, 1991.
- [7] Ciula, Arianna e Eide, Øvind. «Reflections on cultural heritage and digital humanities: modelling in practice and theory». In *First International Conference on Digital Access to Textual Cultural Heritage (DATeCH)*. Madrid, Spagna, 2014. *Conference Proceedings*, pp. 35-41. <http://modellingdh.uni-koeln.de/wp-content/uploads/2017/01/modelling-inpractice-and-theory.pdf>
- [8] Cummings, James. «The Text Encoding Initiative and the Study of Literature». In R. Siemens e S. Schreibman (a cura di), *A Companion to Digital Literary Studies*. Blackwell. Oxford: Blackwell, 2008, pp. 218-239.
- [9] Einaudi, Giulio. *Tutti i nostri mercoledì*. Bellinzona: Casagrande, 2001.
- [10] Ferretti, Gian Carlo. *Storia dell'editoria letteraria in Italia. 1945-2003*. Torino: Einaudi, 2004.
- [11] Flanders, Julia e Jannidis, Fotis (a cura di). *The Shape of Data in Digital Humanities. Modeling Texts and Text-based Resources*. London: Routledge, 2018.
- [12] Iannuzzi, Giulia. «Natalia Ginzburg. Una linea editorial-creativa». In G. C. Ferretti (a cura di), *Protagonisti nell'ombra*. Milano: Unicopli, 2012, pp. 115-133.
- [13] Mangoni, Luisa. *Pensare i libri: la casa editrice Einaudi dagli anni Trenta agli anni Sessanta*. Torino: Bollati Boringhieri, 1999.
- [14] McCarty, Willard. «Modeling: A Study in Words and Meanings». In S. Schreibman, R. Siemens e J. Unsworth (a cura di), *A Companion to Digital Humanities*. Oxford: Blackwell, 2004, pp. 254-270.
- [15] McCarty, Willard. *Humanities computing*. London: Palgrave Macmillan, 2005.
- [16] McCarty, Willard. «Knowing...: Modeling in Literary Studies». In R. Siemens, *A companion to digital literary studies*, Malden: Blackwell 2007, pp. 391-401.
- [17] McGann, Jerome. «Electronic Archives and Critical Editing». *Literature Compass*, n. 7/2 (2010), pp. 37-42.
- [18] Munari, Tommaso (a cura di). *Centolettori: i pareri di lettura dei consulenti Einaudi 1941-1991*. Torino: Einaudi, 2015.
- [19] Munari, Tommaso (a cura di). *I verbali del mercoledì. Riunioni editoriali Einaudi 1953-1963*. Torino: Einaudi, 2013.
- [20] Munari, Tommaso (a cura di). *I verbali del mercoledì. Riunioni editoriali Einaudi 1943- 1952*. Torino: Einaudi, 2011.
- [21] Pierazzo, Elena. «How subjective is your model». In J. Flanders e F. Jannidis (a cura di), *The shape of data in Digital Humanities*. Londra: Routledge, 2018, pp. 117-132.
- [22] Pierazzo, Elena. *Digital scholarly editing. Theories, Models and Methods*. Farnham: Ashgate, 2015.
- [23] Pierazzo, Elena e Leclerc, Élise. «L'edizione scientifica al tempo dell'editoria digitale», *Ecdotica*, n. 12 (2015), pp. 180-193.
- [24] Price, Kenneth M. 2009. «Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?». *Digital Humanities Quarterly*, n. 3/3, <http://digitalhumanities.org:8080/dhq/vol/3/3/000053/000053.html>
- [25] Saita, Nicoletta. «Natalia Ginzburg la fedeltà di una vita con “passo da soldato”». In R. Cicala e V. La Mendola (a cura di), *Libri e scrittori di via Biancamano. Casi editoriali in 75 anni di Einaudi*. Milano: EDUCatt, 2009, pp. 95-124.
- [26] Shillingsburg, Peter L. «How Literary Works Exist: Convenient Scholarly Editions», *Digital Humanities Quarterly*, 3/3 (2009). <http://digitalhumanities.org:8080/dhq/vol/3/3/000053/000053.html>
- [27] Soddu, Paolo (a cura di). *Giulio Einaudi nell'editoria di cultura del Novecento italiano*. Firenze: Olschki, 2015.
- [28] Tomasi, Francesca. «Edizioni o archivi digitali? Knowledge sites e apporti disciplinari», in P. Italia e C. Bonsi (a cura di), *Edizioni critiche digitali: edizioni a confronto*, Roma, Sapienza Università editrice, 2016, 129-136.
- [29] Tortorelli, Gianfranco (a cura di). *Gli archivi degli editori: studi e prospettive di ricerca*. Bologna: Pàtron, 1998.
- [30] Turi, Gabriele. *Casa Einaudi: libri uomini idee oltre il fascismo*. Bologna: Il Mulino, 1990.

# Exploring iconographical and iconological content in semantic data through art historians' interpretations

Sofia Baroncini<sup>1</sup>, Marilena Daquino<sup>2</sup>, Francesca Tomasi<sup>3</sup>

<sup>1</sup> University of Bologna, Italy – sofia.baroncini4@unibo.it

<sup>2</sup> University of Bologna, Italy – marilena.daquino2@unibo.it

<sup>3</sup> University of Bologna, Italy – francesca.tomasi@unibo.it

## ABSTRACT

Iconology and Iconography are branches of art history studies which focus on the recognition and interpretation of the subject matter represented in visual artworks and of the deeper meanings eventually conveyed. Due to its complex and interpretative nature, iconographical and iconological knowledge is usually recorded in catalogue entries in free text fields according to cataloguing standards (e.g. CDWA), making information retrieval challenging. Nevertheless, having such curated data described with the accuracy and flexibility offered by ontologies would allow us not only to newly explore the art history scholarly discourse through quantitative analysis but also to enable a new way to access cultural heritage objects described in the network of Linked Open Data (LOD) through the narratives of experts' interpretations. The current study aims at filling this gap by presenting an RDF dataset on iconographical and iconological interpretations, extracted from the art historian's Erwin Panofsky references and represented according to standards and a new domain ontology. We also present a quantitative analysis and a browsing interface to show the potential of applications for information retrieval and user-centric exploration of curated, domain-specific data.

## KEYWORDS

Data exploration, dynamic digital collections, iconography, iconology, art history, art interpretation, Linked Open Data

## 1. INTRODUCTION AND MOTIVATION

Ontologies are a means for expressing complex semantic relations between objects and creating highly detailed, interconnected digital collections. Therefore, they are a feasible tool for the representation of art historical information about artworks, which often deal with complex information and with art experts' interpretations. In particular, iconographical and iconological content, which concerns the description and interpretation of the subject matter, is characterised by great complexity. In fact, it relates 1) the subjects depicted in a single artwork among each other (e.g. the action of Time who lifts the personification of the Truth), to formal characteristics (e.g. style, perspective), and/or to deeper meanings, 2) the depicted scene to external sources, such as mythological stories or other artworks, 3) the subject with its representative variations of form or meaning over time and place, and 4) the relationship with the context (e.g. religious or cultural practices of the time the artwork was created)[1]. All of these aspects are prone to a degree of subjectivity in the interpretation. Due to its complex nature, iconographical and iconological knowledge in catalogue entries is usually expressed in free text fields according to standards<sup>1</sup>, making thematic information retrieval challenging. Nevertheless, having such curated data described with the accuracy and flexibility offered by ontologies would allow us not only to newly explore the art history scholarly discourse through quantitative analysis but also to furnish a new way to access cultural heritage objects described in the network of Linked Open Data (LOD) through the narratives of experts' interpretations. Our research questions can be expressed as follows:

- How can an ontological modelling of iconographical and iconological interpretations represent the domain features to foster the access, analysis and retrieval of iconographical and iconological content?
- What would be the advantages for users of browsing a semantic network of iconographical and iconological interpretations?

To answer these questions, we focused on a case study of a selection of interpretations made by the art historian Erwin Panofsky. Panofsky was chosen as he is a reference point for the theorization of the iconographical and iconological

---

<sup>1</sup> See CDWA' guidelines for subject matter at [https://www.getty.edu/research/publications/electronic\\_publications/cdwa/18subject.html](https://www.getty.edu/research/publications/electronic_publications/cdwa/18subject.html)



interpretation act [10]. Hence, an ontology (ICON<sup>2</sup>)[20] based on his theory [16] was created, according to which a corpus of interpretations manually extracted from a selection of his books [14, 15, 16, 17] was described in compliance with the RDF standards<sup>3</sup>. Finally, a web application<sup>4</sup> presenting an Exploratory Data Analysis and a thematic gallery of artworks was realised.

## 2. THEORETICAL BACKGROUND

Whereas the term iconography refers to that branch of art history studying the artworks subjects (i.e. iconographies), their attributes, meaning and evolution over time, the current meaning of the term iconology has its roots in the research activity of Aby Warburg [11, 22]. His approach considered the content and forms of the artworks as witnesses of social memory, conducting his analysis in an interdisciplinary way to include religion, culture, and the recurrence of visual patterns through different ages [19, 22]. The first holistic attempt of defining a theory of the iconographical and iconological method was made the Warburg's scholar Erwin Panofsky [11], which remains nowadays a reference point<sup>5</sup>. We refer to [1] for a comparison of Panofsky's theory to the theoretical attempts made by other art historians about the subdivision into levels of the interpretation act.

According to Panofsky, there are different types of meaning that can be interpreted in an artwork, subdivided into three layers. The depth to which the artwork can be understood depends on the background knowledge of the observer, going from a more superficial understanding to a deeper, cultural-related one. In the first level (Pre-iconographical description) natural elements (people, objects, actions) and expressional qualities (emotions) are identified. If the observer has sufficient background knowledge about the subject types and themes that may be represented in the period considered, then he can interpret them as second-level subjects (e.g. a woman is recognised as Venus). Finally, if he/she is aware of the sociocultural context or artist's personality, he can recognise symptoms of such topics in the artwork under examination [16].

## 3. STATE OF THE ART

Currently, several ontologies, vocabularies and knowledge graphs describing art-related topics are available. Among them, CIDOC-crm is the standard for describing cultural heritage (CH) objects [6]. Since it doesn't address domain-specific modelling, it has been expanded by VIR ontology to include iconographical content, the act of interpretation and additional information about it [3]. A first attempt to extend VIR in order to include iconological interpretations has been conducted by [1]. Other related ontologies are HiCO<sup>6</sup>, allowing express interpretation acts in relation to their context, and Simulation ontology [21], which concerns the description of symbols and their symbolic meanings. Besides ontologies, fundamental for expressing complex relations, controlled vocabularies are essential tools for information retrieval. Getty Vocabularies<sup>7</sup> (in particular AAT and IA) and Iconclass<sup>8</sup> provide identifiers for expressing iconographical subjects and terms for describing what is represented by an artwork. The Warburg Institute Iconographic Database is the online resource providing the most complete corpus of images related to the discipline. Although it follows an iconographic index, the categories designed by the historian to classify images often result in blurred groupings that are difficult to understand or replicate without knowing the underlying iconological studies on which they are based.

Among available Knowledge Graphs (KGs), artwork subject matter description is afforded both by domain-specific and generic datasets. [2] Provides an overview of the extent to which iconographical and iconological content is available in RDF datasets accessible through an online SPARQL endpoint, showing that this content is generally poorly represented and limited to a generic subject identification. Another domain KG not included in the study, ArtGraph<sup>9</sup>[4], is in line with the study results. It's also worth mentioning HyperReal[20], a KG expressing an encyclopedic knowledge about symbols and the corresponding symbolical meaning in specific contexts.

The semantic access to online cultural heritage data is a central topic for LOD for CH [7, 8]. As emerges from the literature, the idea of exploiting ontologies to query databases is an established practice [12, 13]. The potential of the application to the cultural heritage domain has been underlined by [22].

Some interesting displays for artwork aggregation were considered also by projects applying Computer Vision (CV) to iconography or feature analysis. Pilka [18] proposes an online tool for creating thematic collections automatically generated

---

<sup>2</sup> Documentation available at (link hidden to preserve anonymity).

<sup>3</sup> RDF data available at (link hidden to preserve anonymity).

<sup>4</sup> <https://iconology-dataset.streamlit.app/>

<sup>5</sup> the main cataloguing standards (CDWA, CCO) for subject matter description in cultural objects refer to this theory.

<sup>6</sup> <https://marilenadaquino.github.io/hico/>

<sup>7</sup> <https://www.getty.edu/research/tools/vocabularies/>

<sup>8</sup> <https://iconclass.org/>

<sup>9</sup> <https://zenodo.org/record/6337958>



on the basis of the user selection<sup>10</sup>, whereas [9] clusters artworks' images according to gesture similarities. Due to the lack of data about iconographical and iconological interpretations, browsing and exploratory functionalities of these themes is, to the best of the authors' knowledge, not available.

#### 4. METHODOLOGY

In order to define the characteristics of the domain, a survey of existing theoretical approaches was conducted, and theories were compared[1]. On the basis of that, a preliminary study on the modelling of 11 case studies grouped in a typological selection from Medieval and Renaissance art was conducted [1]. The study showed that the current ontologies lack means for describing accurately both first-level and deeper meanings of interpretation. This motivated the creation of ICON, an ontology for expressing iconographical and iconological interpretations with a high level of granularity, based on Panofsky's theory of the three-levels interpretation act (cf. section 2) and tested on a corpus of the art historian's interpretations.

The data modelling was realised according to standard ontologies (CIDOC-crm<sup>11</sup>, CiTO<sup>12</sup>, PROV-O<sup>13</sup>) and ICON, following the research questions relevant to the domain defined through a generalization of RQ formulated for the 11 case studies analysed in [1]. The art historian's claims were then described manually according to the model, converted to RDF, and aligned to existing sources.

The identified research questions were used as a guideline to perform exploratory data analysis. As a result, thematic groups of RQs were identified according to the aspect under investigation, namely: cultural phenomena, iconographies and their attributes, symbols, citations of visual motifs, citations of evidence on which the interpretation is based, artwork metadata, different interpretations of the same artwork, and interpretations supporting other ones. The analysis aimed at 1) addressing in a quantitative fashion iconological research questions, and 2) characterising the art historian's approach, examining features of his interpretations, namely: the extent to which a) he makes use of textual sources, b) he describes the artworks at all levels of interpretation, and c) verifying if the complexity of connections embedded in an iconological interpretation emerges from data.

The quantitative overview of the RQs results was, in some cases, integrated with a qualitative insight to allow users to better understand the results and use them as the basis for further research.

#### 5. RESULTS

The created dataset contains interpretations about ca. 400 artworks (see Figure 1) mostly from the Middle Ages and Renaissance Western art, mainly interpreted by Panofsky. The interpretations are divided into three levels, from a more superficial understanding to a deeper one, as described by the art historian's theory, and inter-level links among identified subjects are provided. The subject types recorded include natural elements, actions and emotions (level 1), characters, events, places, objects with a specific identity (e.g. the Bible), personifications, symbols, stories and allegories (level 2), concepts, and cultural phenomena (level 3). For each subject identification a provenance of the assertion can be provided, indicating the author, source, and cited evidence, so as to allow the coexistence of multiple (diverging) interpretations.

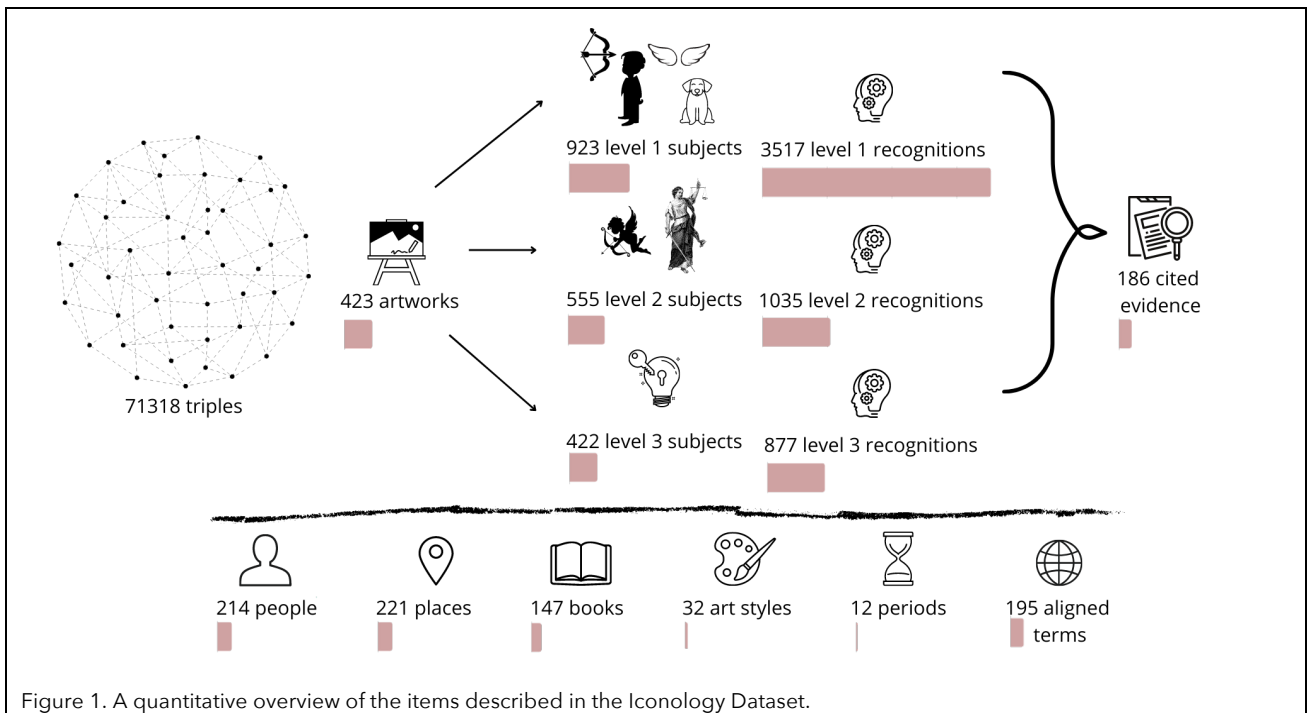
---

<sup>10</sup> <https://digitalcurator.art/>

<sup>11</sup> <https://www.cidoc-crm.org/>

<sup>12</sup> <https://sparontologies.github.io/cito/current/cito.html>

<sup>13</sup> <http://www.w3.org/TR/prov-o/>



The data analysis<sup>14</sup> shows that the art historian focuses on the classical themes, as half of the artworks are involved in the cultural phenomenon “*reception of classical antiquity*” and the most frequent level 2 subjects are related to Classical mythology. The most interlinked types of subjects are natural elements, characters, and cultural phenomena. While it was possible to address in a quantitative fashion all the RQs, some of them were only partially addressed due to the lack of data. In detail, diachronic overviews of the evolution of certain subjects showed limited results, since most subjects rarely appear more than once in the dataset and only 68% of artworks described have a date associated.

Unexpected results were retrieved from the analysis of the art historian’s approach. Despite it has been claimed Panofsky’s method highly relies on textual sources, only 27% of his interpretations cite textual evidence. Instead, artworks are the most cited evidence in iconological recognitions, which is also the type of recognition making more use of evidence, in contrast to pre-iconographical and iconographical recognitions, where texts are the most cited sources. In addition, only 9% of recognitions about artworks part of a book (e.g. illuminations) cite the book’s text as evidence.

Only 53% of artworks are described at all levels, while a high number of artworks have only one level described (61 out of 423), i.e. the third level.

The network of iconological recognitions includes links between recognitions supporting each other and artworks cited as evidence, in turn, related to cultural phenomena. The network graph shows the complexity of iconological interpretations and highlights connections between artworks and phenomena that were originally treated in separate books, otherwise difficult to detect without a visual aid.

Access points for browsing are then provided, namely: the variation of iconographies (i.e. from what level 1 subjects is a level 2 subject composed) providing all the artworks showing a certain variation, the comparison between the artworks in which there is a visual motif borrowing and a gallery of artworks that can be filtered according to a selection of subject types in different levels.

## 6. CONCLUSION AND FUTURE WORK

The presented study provides manually curated domain-specific data based on the authoritative research of the art historian Erwin Panofsky. The data are modelled in the semantic web standard RDF according to ontological standards and a newly created ontology for iconographical and iconological interpretations. As a result, the data provides information about how subjects and meaning are represented in the depicted artworks according to the interpretation, and metadata about the assertion, including evidence, reference, and supportive citations. Therefore, several domain-specific research questions could be answered in a quantitative way. Additionally, information retrieval about iconographies is fostered thanks to their granular description in semantic data and to their linking to controlled vocabularies.

<sup>14</sup> Results and visualisations of the data analysis reported in this section are available at <https://iconology-dataset.streamlit.app/>

The availability of data extracted by the scientific literature of an established art historian fosters the quality of data in the LOD cloud and can be used as a narrative to explore the connected resources in the semantic network. This study is a first case study towards the representation of the scholarly art historian's discourse, currently expressed in natural language in articles and catalogue entries.

Since the dataset includes only the artworks described by the art historian in the selected references, it is not possible to perform a quantitative analysis representative of a certain time period or historical movement. Nevertheless, it can give insights into the art historian's selection and practice. Therefore, future work includes a better alignment with art history KGs representative of the period, such as Zeri&Lode [5], to better study the art historian's personal selection in comparison to the known artworks from a certain period (e.g. Renaissance).

Through the implemented browsing functions, the digital space becomes a dynamic collection of artworks that can be created according to the users' interest in the range of themes and subjects touched by the domain-specific research of iconography and iconology. Future work includes user-centred navigation of the graph, to make the exploration more accessible.

## REFERENCES

- [1] Baroncini, S., Daquino, M., & Tomasi, F. (2021). Modelling Art Interpretation and Meaning. A Data Model for Describing Iconology and Iconography. *AIDAinformazioni*, 1–2, 39–62. <https://doi.org/10.48550/arXiv.2106.12967>
- [2] Baroncini, S., Sartini, B., Van Erp, M., Tomasi, F., & Gangemi, A. (2023). Is dc:subject enough? A landscape on iconography and iconology statements of knowledge graphs in the semantic web. *Journal of Documentation*, 79(7), 115–136. <https://doi.org/10.1108/JD-09-2022-0207>
- [3] Carboni, Nicola and Livio de Luca. 'An Ontological Approach to the Description of Visual and Iconographical Representations'. *Heritage* 2, no. 2 (2019): 1191–1210. <https://doi.org/10.3390/heritage2020078>
- [4] Castellano, Giovanna, Giovanni Sansaro, and Gennaro Vessio. "Integrating Contextual Knowledge to Visual Features for Fine Art Classification," 2021. <https://doi.org/10.48550/ARXIV.2105.15028>
- [5] Daquino, Marilena, Francesca Mambelli, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. "Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data." *Journal on Computing and Cultural Heritage* 10, no. 4 (October 26, 2017): 1–21. <https://doi.org/10.1145/3051487>
- [6] Doerr, Martin. 'The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata'. *AI Magazine* 24, no. 3 (15 September 2003): 75–92. <https://doi.org/10.1609/aimag.v24i3.1720>
- [7] Hyvönen, Eero. 'Semantic Portals for Cultural Heritage'. In *Handbook on Ontologies*, edited by Steffen Staab and Rudi Studer, 757–78. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. [https://doi.org/10.1007/978-3-540-92673-3\\_34](https://doi.org/10.1007/978-3-540-92673-3_34)
- [8] Hyvönen, Eero. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Synthesis Lectures on Data, Semantics, and Knowledge. Cham: Springer International Publishing, 2012. <https://doi.org/10.1007/978-3-031-79438-4>
- [9] Impett, Leonardo Laurence. *Analyzing Gesture in Digital Art History*. The Routledge Companion to Digital Humanities and Art History. Routledge, 2020. <https://doi.org/10.4324/9780429505188-33>
- [10] Liepe, Lena Eva. *The Locus of Meaning in Medieval Art: Iconography, Iconology, and Interpreting the Visual Imagery of the Middle Ages*. Kalamazoo: Medieval Institute Publications, 2019. [https://muse.jhu.edu/pub/241/edited\\_volume/book/68039](https://muse.jhu.edu/pub/241/edited_volume/book/68039)
- [11] Müller, Marion G. 'Iconography and Iconology as a Visual Method and Approach.' In *The SAGE Handbook of Visual Research Methods*, by Eric Margolis and Luc Pauwels, 283–97. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2011. <https://doi.org/10.4135/9781446268278.n15>
- [12] Munir, K., M. Odeh, and R. McClatchey. 'Ontology-Driven Relational Query Formulation Using the Semantic and Assertional Capabilities of OWL-DL'. *Knowledge-Based Systems* 35 (1 November 2012): 144–59. <https://doi.org/10.1016/j.knosys.2012.04.020>
- [13] Munir, Kamran, and M. Sheraz Anjum. 'The Use of Ontologies for Effective Knowledge Modelling and Information Retrieval'. *Applied Computing and Informatics* 14, no. 2 (1 July 2018): 116–26. <https://doi.org/10.1016/j.aci.2017.07.003>
- [14] Panofsky, Erwin. *Meaning in the Visual Arts*. Garden City, NY: Doubleday, 1955.
- [15] Panofsky, Erwin. *Renaissance and Renascences in Western Art*. New York: Harper & Row, 1972. <https://www.worldcat.org/title/1091911205>
- [16] Panofsky, Erwin. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. Boulder, Colo.: Westview Press, 1972.
- [17] Panofsky, Erwin, and Fritz Saxl. 'Classical Mythology in Mediaeval Art'. *Metropolitan Museum Studies* 4, no. 2 (1933): 228–80. <https://doi.org/10.2307/1522803>.
- [18] Pilka, Lukas. "The Cloud of Symbols: Using Computer Vision for Iconographical Analysis of Classical Central European Art." In *Digital Art History—Methods, Practices, Epistemologies: Book Of Abstracts & Conference Programme*. Zagreb, Croatia, 2021. [https://www.ipu.hr/content/info/IPU-DAH-III-2021\\_Book-of-Abstracts.pdf](https://www.ipu.hr/content/info/IPU-DAH-III-2021_Book-of-Abstracts.pdf)
- [19] Rossi Pinelli, Orietta. *La storia delle storie dell'arte*. Torino: Einaudi, 2014.

- [20] Sartini, B., Baroncini, S., van Erp, M., Tomasi, F., & Gangemi, A. (2023). ICON: An Ontology for Comprehensive Artistic Interpretations. *Journal on Computing and Cultural Heritage*. <https://doi.org/10.1145/3594724>
- [20] Sartini, Bruno, Marieke van Erp, and Aldo Gangemi. "Marriage Is a Peach and a Chalice: Modelling Cultural Symbolism on the Semantic Web." In *Proceedings of the 11th on Knowledge Capture Conference, 201–8*. Virtual Event USA: ACM, 2021. <https://doi.org/10.1145/3460210.3493552>
- [21] Tomasi, Francesca, Fabio Ciotti, Marilena Daquino, and Maurizio Lana. 'Using Ontologies as a Faceted Browsing for Heterogeneous Cultural Heritage Collections', Vol. 1st AI\*IA Workshop on Intelligent Techniques At Libraries and Archives. CEUR Workshop Proceedings, 2015.
- [22] Warburg, Aby. *The renewal of pagan antiquity: contributions to the cultural history of the European Renaissance*. Los Angeles, CA: Getty Research Institute for the History of Art and the Humanities, 1999.

# L'applicazione del riconoscimento testi neurale per la realizzazione di ristampe digitali

Elisa Bastianello<sup>1</sup>, Reto Baumgartner<sup>2</sup>

<sup>1</sup> Bibliotheca Hertziana – Istituto Max Planck per la storia dell'arte, Germania/Italia – [elisa.bastianello@biblhertz.it](mailto:elisa.bastianello@biblhertz.it)

<sup>2</sup> Universität Zürich, Svizzera – [reto.baumgartner@zi.uzh.ch](mailto:reto.baumgartner@zi.uzh.ch)

## ABSTRACT

Con il miglioramento delle tecnologie per il riconoscimento dei testi, sono finalmente disponibili delle trascrizioni automatiche sempre più accurate. Eppure queste trascrizioni, quando sono trasformate in testo modificabile, richiedono spesso un grandissimo lavoro di post produzione, in particolare per l'identificazione della struttura del testo, i titoli, le note a piè di pagina, o la formattazione speciale dei caratteri, come i corsivi, gli apici, i maiuscolotti. Questi testi, che funzionano abbastanza bene in caso di ricerca full text, non sono invece sufficienti quando lo scopo del riconoscimento è una edizione o una ristampa digitale, in particolare in formato TEI XML.

In questo articolo viene illustrato un workflow sperimentale messo a punto per provare ad ovviare queste limitazioni con l'ausilio del riconoscimento neurale del testo e del layout di pagina attraverso la piattaforma Transkribus, e successivo trattamento del risultato attraverso un processo di sostituzione con espressioni regolari e XSLT, fino alla pubblicazione del risultato sulla piattaforma TEI Publisher per il successivo arricchimento.

## PAROLE CHIAVE

Trascrizione neurale; Ristampa digitale; TEI XML; Transkribus

## 1. 1. INTRODUZIONE

La scansione digitale dei documenti è stata accompagnata già agli inizi del secolo XX da sistemi di conversione dei caratteri, soprattutto ad uso delle persone cieche o con limitazioni alla vista. La tecnologia dell'OCR (Optical Character Recognition) ha ricevuto una spinta avanti molto forte dall'introduzione, alla fine del secolo, di tecnologie open source disponibili nel web, con accuratezza sempre maggiore delle trascrizioni, grazie al miglioramento delle tecniche di digitalizzazione dei documenti e dall'integrazione con vocabolari.

In ambito librario le trascrizioni ottenute dai software di OCR sono utilizzate per fare ricerca a testo intero (per esempio su Google Books), o per fornire versioni PDF ricercabili, dove insieme all'immagine scansionata si trova un livello di testo semplice posizionato secondo le coordinate dei caratteri trascritti, oltre a permettere l'uso di strumenti TTS (Text to Speech) per aumentare l'accessibilità.

Quando però da questo testo riconosciuto vogliamo ricavare una trascrizione da riusare in modo nativamente digitale, per esempio citare un paragrafo all'interno di un saggio senza trascriverlo manualmente, ci accorgiamo subito che, in alcuni casi, la conversione non è ottimizzata: i paragrafi sono spezzati in singole righe, elementi come titoli correnti o numeri di pagina possono essere inseriti all'interno del testo principale. Difficilmente inoltre il testo trascritto conserva correttamente la formattazione dei caratteri, come le parole in corsivo o i riferimenti alle note a piè di pagina in apice. La situazione si complica nel caso in cui, anziché poche righe, andiamo a riprodurre l'intero testo di un libro, in particolare di un libro scientifico con una struttura di paragrafi, sottoparagrafi, note e citazioni nel testo. Se poi consideriamo la trascrizione di testi la cui ortografia non è normalizzata, come il vernacolo italiano, o testi a stampa antichi, con legature, abbreviature, forme alternate come la *f* (s lunga) o font di difficile decodifica come quelli gotici, anche l'accuratezza della trascrizione subisce un sensibile ribasso per l'impossibilità di utilizzare un vocabolario di riferimento.

Questo significa che per una ristampa digitale a partire da scansioni di libri esistenti si rende necessaria non solo una accurata revisione del testo alla ricerca di errori di riconoscimento (per esempio lo scambio del numero 1 con la lettera l o I maiuscola), ma anche una totale ricostruzione della struttura del testo in paragrafi, riconnessione delle note ai rispettivi riferimenti e l'attribuzione degli stili di carattere (o in formato XML dei tag di rendition) corretti. Sebbene esistano dei servizi esterni specializzati, il costo di tale conversione risulta estremamente oneroso sia in termini di costi che di tempo. Questo progetto è un primo tentativo di sfruttare le macchine neurali e creare un flusso di lavoro integrato dal costo limitato e dai risultati configurabili in base all'uso finale del testo [1, 6, 7].

## 2. IL LIBRO CAMPIONE

*Raphael in Early Modern Sources, 1483–1602* di Jonh Shearman [5] è un'opera composta da due volumi per un totale di quasi 1700 pagine di testo, e raccoglie la trascrizione di 1.058 documenti datati tra il 1483 ed il 1602 relativi alla vita e alle

opere dell'artista Raffaello Sanzio, insieme a 58 documenti "falsi" ma importanti per la storiografia dell'urbinate. Il testo, nella sua forma cartacea, ha rappresentato una pietra miliare per la ricerca e già da alcuni anni risulta fuori commercio e molto ricercato tra i libri usati. La proposta di una ristampa in formato digitale andava dunque non solo a sopperire ad una reale richiesta da parte degli studiosi, ma anche a moltiplicare le modalità di accesso, annotazione e studio dei documenti stessi e degli apparati critici, soprattutto in un momento come l'inizio del lockdown del 2020, quando le biblioteche pubbliche avevano dovuto contingentare l'accesso diretto ai libri. Purtroppo le ricerche dei file originali negli archivi Yale University Press e della Bibliotheca Hertziana – Istituto Max Planck per la storia dell'arte (che hanno pubblicato il libro in origine), come anche in quelli dell'autore non ha avuto gli esiti sperati e non è stato possibile rinvenire documenti in formato digitale nativo. È stato perciò necessario procedere alla scansione dei due volumi<sup>1</sup> per poi riconoscere il testo e convertirlo in formato TEI XML.

Il contenuto del testo presenta esattamente le criticità che rendono la conversione con tecnologie OCR tradizionali estremamente complessa e di conseguenza esosa sia in termini economici che di tempo:

- Ad esclusione dell'introduzione critica iniziale (pagine xi-41), e dei commenti dell'autore ai documenti, in lingua inglese corrente, il contenuto è costituito da trascrizioni di documenti d'epoca, per lo più in vernacolo italiano e latino rinascimentale, indici e referenze bibliografiche, che non si prestano al supporto di vocabolari per il miglioramento del riconoscimento.
- La struttura della sezione documenti è molto articolata: insieme ai titoli possono comparire identificativi, abstract, corpo della trascrizione, nota sulla fonte, commento e bibliografia. Per questo in ogni pagina sono presenti diverse dimensioni del font, con corsivi, maiuscolotti, grassetto e apici. Utilizzare un normale OCR implica un lavoro di revisione enorme su due tomi già complessi per la lunghezza, con un controllo pagina per pagina, come confermato dai servizi esterni contattati per valutare il progetto.

Allo stesso tempo, le dimensioni dei due volumi giustificavano i tempi necessari per la creazione di un processo alternativo basato sull'utilizzo delle macchine neurali per il riconoscimento del testo, inclusi quelli per la creazione dei modelli iniziali, rendendo l'opera il progetto campione ideale con cui valutare i vantaggi e i problemi derivanti da questo processo.

### 3. LA TECNOLOGIA

Da alcuni anni sono state sviluppate tecnologie basate sull'apprendimento automatico (*machine learning*) per il riconoscimento dei testi, in particolare quelli manoscritti, note come HTR (*Handwritten Text Recognition*)<sup>2</sup>, che sono in grado di migliorare l'accuratezza delle trascrizioni ampliando il contesto di riferimento dal singolo carattere all'intera riga [3]. Queste stesse macchine neurali possono essere usate anche sui testi a stampa, e anzi stanno prendendo sempre più piede, al punto che si preferisce definirle motori OCR neurali, per evitare una definizione legata al solo testo manoscritto. L'utilizzo diretto dei motori richiede delle buone conoscenze informatiche, dimestichezza con la linea di comando e server dedicati. Per fortuna per chi non ha a disposizione le infrastrutture e le competenze informatiche necessarie esistono delle piattaforme che offrono l'uso dei motori installati nei loro server attraverso una piattaforma in grado di guidare l'utente in quasi tutte le fasi del processo. In particolare dal 2016 (originariamente grazie ad un finanziamento europeo) è disponibile la piattaforma Transkribus [2]. Per poter sfruttare il supporto dei motori neurali, la piattaforma richiede che tutte le immagini vengano caricate nei loro server remoti, da cui sono poi disponibili attraverso il client desktop, o più recentemente, l'interfaccia web<sup>3</sup>. L'uso di immagini in cloud permette di lavorare in gruppo sullo stesso documento, mentre l'interfaccia grafica consente di avere un controllo completo sia della segmentazione (disegno sul documento delle aree di testo e delle linee) che di trascrivere i contenuti. Lo strumento offre anche l'esportazione dei risultati in vari formati, tra cui TEI XML, che era il formato prescelto per la pubblicazione, sebbene su questo argomento tornerò in seguito per alcune precisazioni.

Il primo passo è stato dunque quello di caricare le scansioni dei due volumi sul server centrale di Transkribus, per poter utilizzare le funzionalità di riconoscimento testo e struttura basate su rete neurale. Per generare una prima trascrizione,

---

<sup>1</sup> Vorrei ringraziare i colleghi Johannes Röhl e Susanne Kubersky per aver suggerito il volume per la ristampa digitale nell'aprile 2020, complice il cinquecentenario della scomparsa dell'Urbinate, e Klaus Werner per avere pazientemente scansionato i due volumi altrimenti inaccessibili a causa delle limitazioni legate alla pandemia COVID19.

<sup>2</sup> Alcuni esempi di motori neurali per il riconoscimento dei testi sono HTR+ sviluppato da CITlab (<https://www.mathematik.uni-rostock.de/en/forschung/projekte/citlab/> - proprietario) e Pylaia (<https://github.com/jpuigcerver/pylaia> - open source).

<sup>3</sup> Recentemente sono state sviluppate altre piattaforme, più orientate all'open source e installabili localmente dall'utente finale, come eScriptorium (<https://www.escriptorium.uk/>), ma nessuna di esse era sufficientemente matura al momento di questo progetto, soprattutto in assenza di infrastrutture per la implementazione e personale per la gestione.



trattandosi di un documento recente a stampa, è stato utilizzato il servizio OCR (ABBYY Fine Reader v. 11) che al tempo della realizzazione era integrato al client<sup>4</sup>. A differenza dall'uso diretto di ABBYY nella sua interfaccia nativa, nel caso sia necessario addestrare un modello di riconoscimento specifico, non è necessario correggere pazientemente pagine e pagine di trascrizione un carattere alla volta. È infatti possibile generare un nuovo modello semplicemente correggendo la trascrizione riga per riga per creare il testo di base (*Ground Truth*, GT) da cui generare un nuovo modello per l'apprendimento automatico. Inoltre, dato che l'area analizzata è l'intera riga e non il singolo carattere, è molto più semplice considerare elementi complessi, come le legature, nella loro interezza e addirittura fare corrispondere più lettere ad un segno. Inoltre sono disponibili numerosi strumenti per l'annotazione, cioè la marcatura, sia della struttura semantica delle aree di testo che del testo stesso.

#### 4. ANNOTAZIONE E SOLUZIONI CREATIVE

Dopo la prima fase il testo era stato segmentato in aree di testo corrispondenti generalmente ai singoli paragrafi, ma mancava un valore strutturale. Per poterlo esportare in TEI era necessario distinguere la struttura, ovvero banalmente, elementi come i titoli correnti, il corpo del testo, i numeri di pagina o le note a piè di pagina. Uno dei vantaggi di Transkribus è che non solo è possibile assegnare una valenza strutturale a tutti i paragrafi con tag con una corrispondenza diretta con la struttura XML TEI da usare in fase di esportazione, ma è anche possibile addestrare un modello perché assegni tali tag in automatico alle aree di testo esistenti grazie alla tecnologia P2PaLA<sup>5</sup>.

Nel caso specifico, a livello del layout di pagina, sono stati marcati tutti i numeri di pagina, i titoli correnti, i titoli di paragrafo e le note a piè di pagina in maniera distinta dal testo nei capitoli introduttivi. Per quanto riguarda gli oltre 1000 documenti trascritti, è stato possibile identificare anche per essi una sequenza comprendente numero del documento, titolo, abstract, trascrizione, nota archivistica, commento e bibliografia, che si ripete, sebbene in modo non sempre completo. Dopo aver annotato manualmente circa 50 pagine di struttura, è stato addestrato un primo modello in grado di assegnare automaticamente buona parte degli attributi strutturali al documento. Questa tecnologia necessita di una chiara distinzione “grafica” tra gli elementi da segmentare o marcare<sup>6</sup>, ovvero non deve dipendere dalla lettura del contenuto ma solo dal “colpo d’occhio”. Per poter ricostituire i paragrafi che superavano lo spazio della singola pagina (come vedremo più avanti) si è scelto di assegnare l’attributo “continued” a tutti gli elementi che continuavano (per es. footnote/footnote-continued, paragraph/paragraph-continued), ma questa distinzione non era possibile senza leggere il contenuto e per questo è stato necessario revisionare a mano la segmentazione. In futuro l’uso di *Natural Language Processing* potrebbe individuare buona parte dei casi in cui il testo del paragrafo non è completo e continua nel paragrafo successivo o precedente, ma al momento questa strada non è stata presa in considerazione. [fig.1] La correzione manuale si è resa necessaria anche per verificare le aree di testo esistenti<sup>7</sup> non sempre corrispondenti ai singoli paragrafi. Inoltre, anche se sono stati generati nuovi modelli man mano che veniva corretto il testo esistente, quando sono stati aggiunti stili specifici per casi speciali come le iscrizioni, la poesia o il testo centrato, si è rilevato come la presenza di tag poco usati nel modello aumenta il numero di paragrafi marcati in modo erroneo in fase di riconoscimento. Per questa ragione nel modello finale si è preferito limitare il numero dei tag a quelli con più alta frequenza (lo strumento stesso consente di analizzare quantitativamente i tag assegnati nel testo con la funzione “Analyze structure types” da cui si rileva, per esempio che il rapporto tra i paragrafi poetici e i paragrafi di corpo del testo è circa 150 a 3000, cioè 5:100).

---

<sup>4</sup> Dopo la rimozione della funzionalità OCR dalla piattaforma, è possibile ottenere risultati comparabili con il riconoscimento dei blocchi di testo e l’uso di modelli pubblici come Transkribus Print M1 <https://readcoop.eu/it/model/transkribus-print-multi-language-dutch-german-english-finnish-french-swedish-etc/>.

<sup>5</sup> Page to PAGE Layout Analysis, sviluppato dal progetto <https://github.com/lquirosd/P2PaLA>.

<sup>6</sup> È possibile usare P2PaLA per tracciare le aree ex novo (Create new layout), ma nel caso specifico si è usata la funzione di marcatura delle aree esistenti, quelle generate da ABBYY (Label existing transcription).

<sup>7</sup> Vorrei ringraziare Viviana Nocerino, Iolanda Pagano e Andrea Pecorella per il prezioso supporto in questa fase di lavoro.



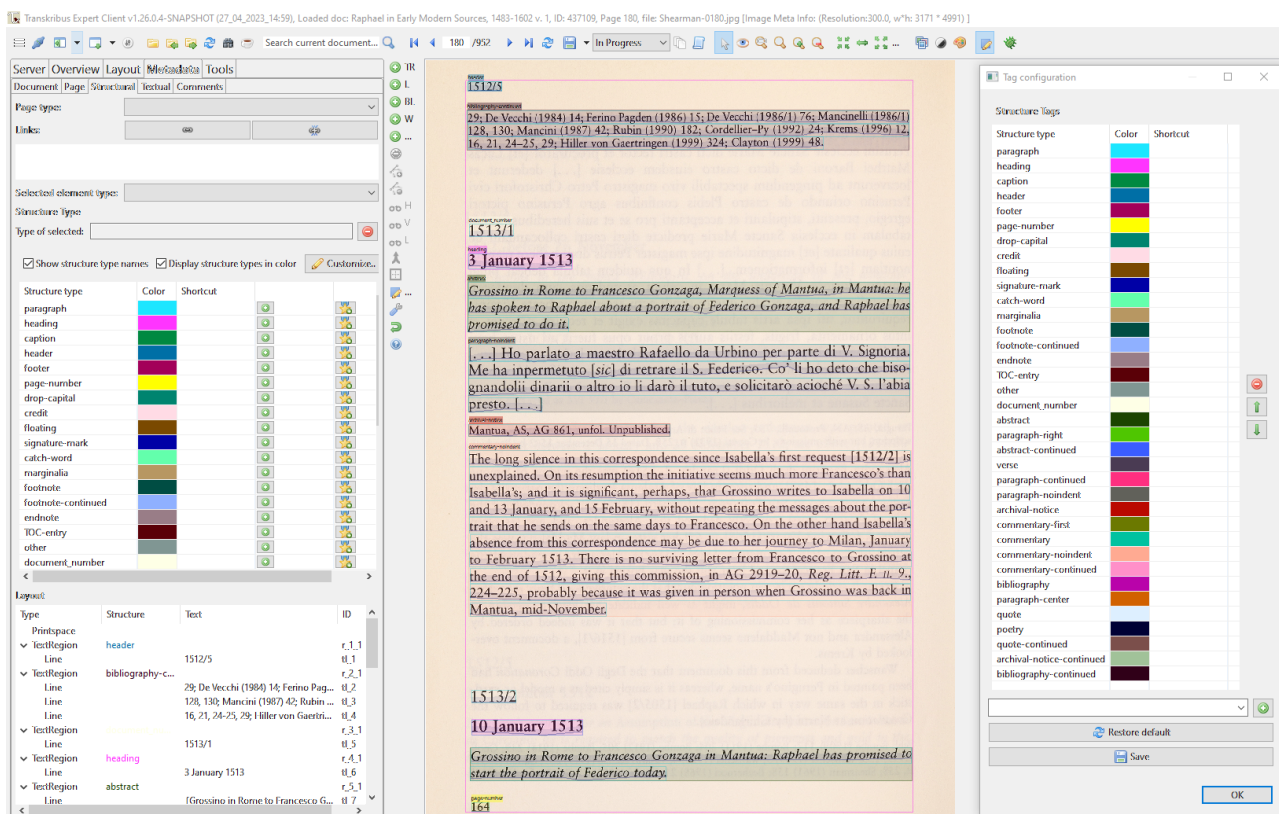


Figura 5. La completa segmentazione strutturale dei documenti trascritti in Raphael in Early Modern Sources.

Per quanto riguarda il testo, la trascrizione ottenuta dall'OCR sarebbe stata già sufficiente per trascrivere il contenuto, ma non conteneva nessuna informazione sulle alterazioni dei caratteri di base, come testo in corsivo, grassetto, maiuscoletto o i numeri in apice di rimandi alle note a piè di pagina. Gli strumenti OCR tradizionali possono in particolari occasioni distinguere alcune alterazioni, ma solo se il modello identifica il singolo carattere come corsivo o grassetto all'interno di un testo il cui corpo è dimensionalmente consistente. In generale questo tipo di trascrizione è poco accurata, con numeri in apice abbassati al livello del testo e corsivi incongruenti. Al contempo le macchine neurali di riconoscimento testo, proprio per il fatto di essere state sviluppate per poter riconoscere manoscritti, non sono predisposte per segnalare questo tipo di informazioni, che vengono aggiunte come annotazioni del contenuto. Questo però non significa che non siano in grado di rilevare le differenze, dato che si tratta di elementi graficamente distinguibili. Serve però un metodo in grado di permettere al modello di restituire anche queste informazioni insieme al testo. La soluzione trovata è quella di addestrare un cosiddetto *smart model*, ovvero un modello neurale che non si limita a trascrivere il testo ma aggiunge informazioni. Il caso più frequente di modello smart è quello che interpreta le abbreviature e restituisce le espansioni a seconda del contesto. Questo significa che per esempio, invece di trascrivere "Vitruuio", con la u consonantica, trascrive normalizzando in "Vitruvio", oppure scioglie "p" con "per", "q" con "quod" e i compendi di m e n (ō, ā ecc.) con la necessaria espansione (on/om, an/am) a seconda del contesto. Nel nostro caso però non si tratta di informazioni testuali, ma di tipo grafico e le informazioni aggiuntive dovevano essere veicolate in una forma diversa. Quando si codifica un testo in formato digitale questo tipo di informazioni sono fornite da "tag", per esempio nell'HTML `<i>...</i>` è il corsivo, `<sup>...</sup>` l'apice e così via. Prendendo spunto da esperienze di altri utenti, che per identificare gli spazi multipli tra parole li indicavano tra parentesi, alla trascrizione del contenuto sono stati aggiunti dei marcatori per segnalare l'inizio e la fine di ciascuna alterazione. Dato che le parentesi, come anche altri caratteri già presenti nel testo avrebbero reso ambigua la loro ricerca, si è preferito l'utilizzo di simboli UNICODE, per la precisione coppie di simboli matematici, univoci nel testo e quindi facilmente sostituibili in fase di post produzione. Per esempio sono stati utilizzati:

- $\int$  (U-2320/U+2321) per indicare il corsivo
- $\subset\supset$  (U+2282/U+2283) per indicare il testo in apice nei rimandi alle note nel testo
- $\top\perp$  (U+22A4/U+22A5) per indicare il testo in apice nei riferimenti delle note a piè di pagina
- $\textcircled{0}$  (U+208D/U+208E) per indicare il maiuscoletto
- $\mathbb{N}$  (U+21BE/U+21C3) per indicare il grassetto

L'unica accortezza è stata quella di marcare manualmente prima e dopo ogni parola o gruppo di parole per ogni riga, dato che un primo modello, in cui le frasi in corsivo che andavano a capo erano state marcate solo all'inizio e alla fine dell'inciso aveva dimostrato di aver problemi nell'inserire marcature alla fine o all'inizio della riga. A posteriori questo problema è una ovvia conseguenza del fatto che i modelli analizzano le singole righe e non gli insiemi di righe e il loro contenuto. [fig.2] Sono state sufficienti circa 50 pagine annotate a mano per generare un primo modello per il motore CITlab HTR+, con il quale sono state riconosciute ex novo le pagine successive, iterando tra correzione del testo trascritto e generazione di un modello più completo. A differenza del modello di struttura, il modello di riconoscimento del testo è risultato estremamente consistente ed affidabile, anzi, i maggiori problemi sono stati quelli legati agli errori nell'OCR delle pagine di GT iniziale, dato che all'epoca non erano disponibili modelli neurali pubblici di riconoscimento del testo a stampa recente a cui fare riferimento per le lettere mancanti nell'introduzione, per esempio accenti francesi, diresi tedesche, tilde spagnole che comparivano solo a livello di note o bibliografia.

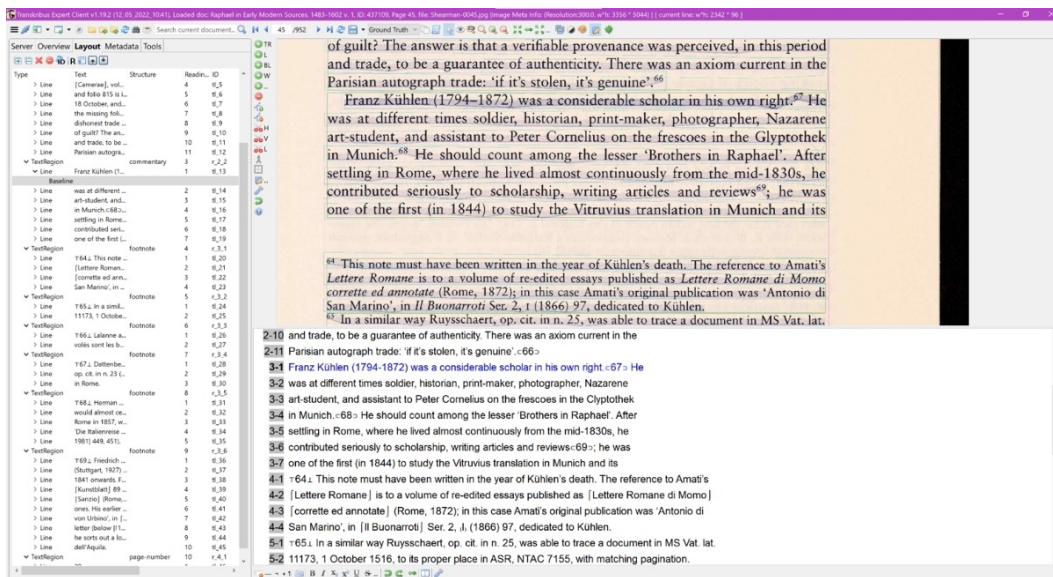


Figura 2: Trascrizione con marcatura degli stili di testo in Raphael in Early Modern Sources.

Anche in questo caso prima della esportazione è stata necessaria una accurata correzione a mano, che nel complesso, tra struttura e contenuti è durata circa tre mesi, includendo anche le bibliografie e gli indici finali.

## 5. ESPORTAZIONE E CONSIDERAZIONI FINALI

Come accennato in precedenza, Transkribus prevede tra le sue esportazioni quella in formato TEI XML, ma si tratta di una trasformazione ottimizzata per manoscritti con facsimile a fronte e non particolarmente indicata nel caso di tag strutturali personalizzati come quelli utilizzati nei due volumi dato che i tag non standard vengono omessi.

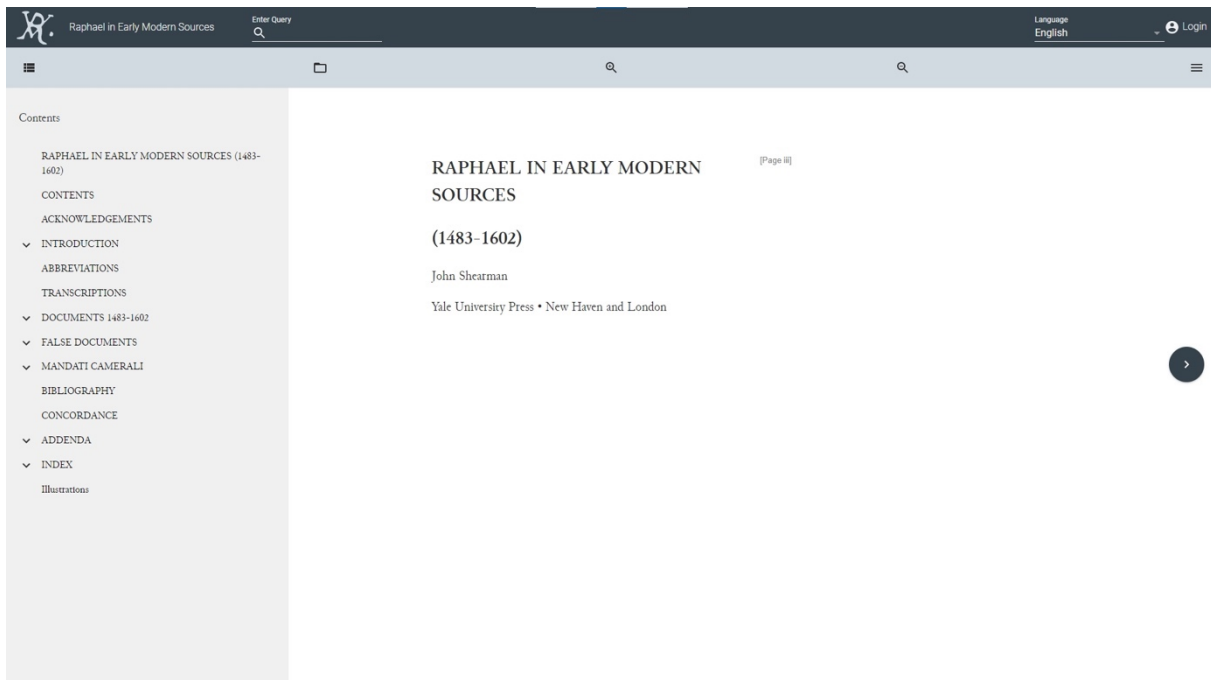
Per questa ragione si è preferito salvare localmente le trascrizioni in formato nativo PAGE XML [4] per convertirle in TEI XML grazie ad una trasformazione XSLT sviluppata modificando PAGE2TEI<sup>8</sup>, in grado di convertire la maggior parte delle aree di testo standard nel corrispettivo elemento strutturale, inclusa la marcatura della paginazione, e mantenendo come “anonymous block” (ab) tutti i tag non standard. La versione del progetto è stata modificata per escludere interamente i facsimile dei volumi originali, non necessari in una ristampa digitale di un testo così recente, e migliorare la conversione dei tag strutturali specifici dei documenti.

I simboli di marcatura delle varianti di aspetto dei caratteri sono stati sostituiti con i rispettivi tag TEI tramite semplici sostituzioni REGEX. Più complessa è stata la ricostituzione del testo, in particolare quello dei paragrafi che continuavano su più pagine e delle note a piè di pagina. L'idea iniziale era quella di marcare in modo distinto i paragrafi che si interrompevano da quelli interamente completi nella pagina e quelli che rappresentavano la continuazione di paragrafi precedenti. Il fallimento dell'addestramento sul riconoscimento di queste varianti e l'idea di analizzare il documento a partire dal fondo verso l'inizio, ha ridotto la distinzione a quella tra paragrafi che continuavano e paragrafi normali, con i primi attaccati ai precedenti paragrafi dello stesso tipo via via fino al raggiungimento di un paragrafo base. Con una serie di trasformazioni XSLT successive e sostituzioni REGEX globali è stato possibile:

<sup>8</sup> Dario Kampkaspar <https://github.com/dariok/page2tei>

- utilizzare il contenuto delle aree di testo marcate come “page\_number” come etichetta delle nuove pagine
- riunire il contenuto delle note a piè di pagina con il rimando nel testo, grazie all’identità tra i numeri in apice marcati nel testo e nelle note
- riunire le parole sillabate nel testo in parole intere. I trattini di a capo nella trascrizione sono marcati con il simbolo ꞑ (U+00AC), ma non volendo rischiare di unire parole normalmente scritte con il trattino con una sostituzione automatica, si è cercato di usare dei vocabolari (non utili per l’italiano antico) per ridurre gli errori.
- collegare i rimandi tra i documenti (indicati come numeri in grassetto nel testo) con i documenti corrispondenti. Anche per questo la presenza di gruppi di documenti nei rimandi ha reso impossibile il totale completamento in modo automatico.

Il primo risultato di questo lavoro è stato il caricamento online della prima versione della ristampa digitale su una istanza di TEI Publisher<sup>9</sup> accessibile all’indirizzo <http://rems.humanitiesconnect.pub>. [fig. 3a-d] Anche se si tratta di una istanza BETA non ancora definitiva, è possibile toccare con mano i vantaggi della segmentazione automatica sia per la presenza degli indici dinamici che durante le ricerche a testo intero. Infatti le ricerche possono essere filtrate tra testo dei commenti e delle note e testo delle trascrizioni, grazie alla configurazione del motore di ricerca della piattaforma eXist-db<sup>10</sup> su cui funziona TEI Publisher.



<sup>9</sup> <https://www.teipublisher.com>. La ristampa digitale è disponibile in formato beta all’indirizzo <http://rems.humanitiesconnect.pub>.

<sup>10</sup> <http://exist-db.org/>, il motore è basato su Apache Lucene <https://exist-db.org/exist/apps/doc/lucene>.

Raphael in Early Modern Sources

Enter Query

Language English

Contents

- RAPHAEL IN EARLY MODERN SOURCES (1483-1602)
- CONTENTS
- ACKNOWLEDGEMENTS
- INTRODUCTION
- Laus Goltzi
- Reflections
- The Resistance to Documents
- Forging Raphael's Curriculum Vitae
- Facsimiles and Their Market
- A Roman Scriptorium
- Epilogue
- ABBREVIATIONS
- TRANSCRIPTIONS
- DOCUMENTS 1483-1602
- FALSE DOCUMENTS
- MANDATI CAMERALI
- BIBLIOGRAPHY
- CONCORDANCE
- ADDENDA
- INDEX
- Illustrations

**Reflections**

Seventy-five years ago Julius von Schlosser grumbled about 'the absolute impotence in the critical and analytical study of sources' in art history, and I have to say, at the end of my study of the critical history of these documents on Raphael, that from this not unfair sample it seems to me that Schlosser was justified and - what is worse - still is. <sup>1</sup> I choose two examples.

The *Breve* of Pope Leo X appointing Raphael *praefectus* of building stone excavation for New Saint Peter's was written by Pietro Bembo, and it exists in *principes* of his papal letters of 1536. Its use in historiography has been subject to two be dismissed by some as being of concern only to pedants, confuses the date, which as 27 August 1515. A survey of about a hundred and fifty citations shows that Reumont and Milanesi, give wrong dates, even by three years in one case. Wrong strings of authors. The other kind of error lies in reading the matter in such a way that it does not: the *praefectus marmorum et lapidum omnium* becomes *Soprintendente* *Denkmalpfleger, Ispettore generale delle belle arti, Oberaufseher der römischen Altertümer Konservator der antiken Monumente der Stadt Rom*. And even when the more scrupulous abuse, their writings are ignored and the abuse continues to this day. This is not a problem doing well in the reading of documents, and we don't read the secondary literature.

A second example: we have needed to know, it seems with increasing anxiety, about the author or authors of the 'programmes' of the Stanze, and it is becoming almost predictable that Paolo Giovio's comment (c. 1525) on (Page 4) the patronage of Julius II will be quoted yet again as if it told us what we want to know: as if it said that Raphael painted to the 'prescription' of Pope Julius. Two errors again: the first lies in translating *ad praesciptum Julii pontificis* as 'to the prescription of Pope Julius' when it means, simply and more neutrally, 'by the command of'. The second error is the not uncommon one - which plagues the interpretation of Michelangelo's poetry too - of not reading the whole text, or in this case of not even reading the whole sentence, which says the following: 'In the Vatican, when his authority was not yet established, he [Raphael] painted two chambers by command of Pope Julius; Giovio's point had nothing to do with iconographic programmes, but with Julius's rather surprising confidence in an untried artist. I don't believe Schlosser would be any more

Julius [von] Schlosser, *Die Kunstliteratur. Ein Handbuch zur Quellenkunde der neueren Kunstgeschichte* (Vienna, 1924), 266-67: 'Von dieser Grundlage ist also auszugehen, wollen wir den Historiker Visconti richtig verstehen und würdigen. Noch der verdienstvolle Milanesi behandelt ihn wie einen modernen Schriftsteller, bemißt Lob und Tadel aus heutigen Ansprüchen und Erfahrungen heraus: das deribte Verkehrtste und ein neuer Beweis für die vollkommenste Hilflosigkeit der kunsthistorischen Disziplin quellenkritischer Betrachtung gegenüber! Von allen historischen Wissenschaften steckt die Kunstgeschichte hier sicher am längsten in den Kinderschuhen [...]. Schlosser's metaphor seems to be borrowed from, and to refer to, Aaron Spitzer's much earlier critical observation of the immaturity of the discipline struggling for recognition (Kunstketneten und Kunsthistoriker', in *Im neuen Reich* 2/2 [1881] 750: 'Das dieses Verhältnis so häufig übersehen, der rein wissenschaftliche Charakter der Kunstgeschichte nicht streng genug festgehalten werde, hängt mit der grossen Jugend des Faches zusammen. Kaum dass die Kunstgeschichte den Kinderschuhen entwachsen ist'. On Spitzer's essay see Kathryn Brush, *The Shaping of Art History. Wilhelm Vöge, Adolph Goldschmidt, and the Study of Medieval Art* (Cambridge, 1996) 23, 186.

https://staging.humanitiesconnect.pub/evist/apps/remis/remis.vml?id=ch-introduction-2Ph\_facs\_v1\_19\_note\_4...e statistics in this case, and I shall not go on in this vein. I don't believe Schlosser would be any more

Raphael in Early Modern Sources

Enter Query

Language English

Contents

- RAPHAEL IN EARLY MODERN SOURCES (1483-1602)
- CONTENTS
- ACKNOWLEDGEMENTS
- INTRODUCTION
- ABBREVIATIONS
- TRANSCRIPTIONS
- DOCUMENTS 1483-1602
- FALSE DOCUMENTS
- MANDATI CAMERALI
- BIBLIOGRAPHY
- CONCORDANCE
- ADDENDA
- INDEX
- Illustrations

**1494/4**

**1 August 1494**

*Death of Raphael's father, Giovanni Santi.*

Al nome di Dio, di 27 di luglio 1494

Convento di San Francesco di dare, per più cere intrati per più morte e ufilie e luminarii, come appare in questo a carta 49

[...]

E a di ditto [primo d'agosto], per la morte di Giovane di Santi - l. 14 carlini 8

[...]

Urbino, Archivio del Convento di San Francesco, *Entrata e Uscita* (1485-97 and 1520-21), fol. 50v; published by Pungileoni (1822) 137.

Note also the addition to the testament of 27 July, above [1494/2]: 'Die prima mensis augusti decessit dicitur restator, et ego [vibe notary, Lodovico di Antonio] interfui eius funeri et sepulture [...]' (Page 61)

Bibl: Pungileoni (1822) 137; Rehberg (1824) 34; Passavant (1839) 1, 408; Rosini (1843) 27; Boucher-Desnoyers (1852) 11; Passavant (1882-) 1, 273; Waischer (1926) 160; Goltzi (1936) 4; Dubos (1971) 31; Henry (1999) 223.



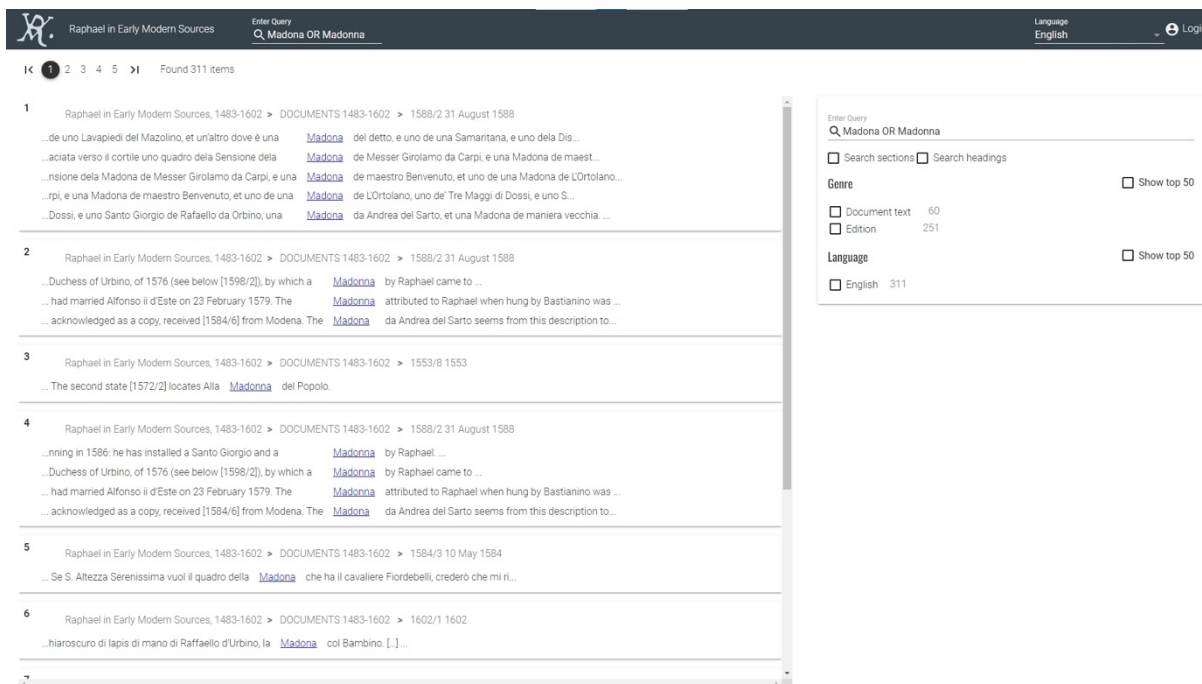


Figure 3a-d. Schermate della ristampa digitale di Raphael in Early Modern Sources: (a) frontespizio e struttura, (b) testo del commento e note, (c) testo del documento, (d) ricerca avanzata.

La stessa piattaforma Transkribus ha preso atto delle possibilità dei modelli smart integrandone il meccanismo direttamente nelle opzioni di addestramento dei modelli. Da qualche tempo infatti è possibile selezionare l'opzione di includere i tag, e in particolare quelli relativi alle alterazioni del testo e alle abbreviature, direttamente nel modello. Nuovi motori di riconoscimento neurale, come Kraken<sup>11</sup> o basati su trasformatori, permetteranno presto di riconoscere direttamente la struttura principale dei testi. Per questa ragione, in collaborazione con il team di sviluppo di Transkribus, stiamo portando avanti un progetto pilota di riconoscimento dei testi a stampa, indipendentemente dalla loro epoca di produzione, e la successiva conversione in formato TEI.

Per quanto riguarda l'edizione di Shearman, i prossimi passi riguarderanno probabilmente l'annotazione della bibliografia, collegando le sigle bibliografiche alla bibliografia estesa, l'integrazione con le pubblicazioni scientifiche più recenti che hanno corretto ed integrato il corpus originale e l'annotazione delle entità (intese come *named entities*) nei contenuti trascritti e nei commenti.

Il maggiore risultato del progetto è che il processo è già stato utilmente riutilizzato per altre trascrizioni digitali per edizioni critiche, dove, con minime alterazioni, è stato applicato a testi dell'inizio del XX secolo all'interno dell'opera completa di Heinrich Wölfflin<sup>12</sup> in sostituzione del normale procedimento che includeva OCR e riedizione manuale.

Condividendo il nostro lavoro speriamo di ispirare anche altri gruppi di lavoro e semplificare l'opera di preparazione della base digitale di edizioni critiche<sup>13</sup>. Siamo consapevoli che alcune delle trasformazioni sono peculiari e specifiche del progetto, ma riteniamo che possano comunque fornire degli spunti per trovare soluzioni diverse e aprire un dialogo sempre maggiore tra quanti si occupano di edizioni e ristampe digitali.

## BIBLIOGRAFIA

- [1] Ryan Cordell. 2017. "Q i-jtb the Raven": Taking Dirty OCR Seriously. *Book History* 20, 1, 188–225. DOI: <https://doi.org/10.1353/bh.2017.0006>
- [2] Philip Kahle, Sebastian Colutto, Gunter Hackl, and Günter Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 19–24. DOI: <https://doi.org/10.1109/ICDAR.2017.307>
- [3] Joe Nockels, Paul Gooding, Sarah Ames, and Melissa Terras. 2022. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Arch Sci* 22, 3, 367–392. DOI: <https://doi.org/10.1007/s10502-022-09397-0>

<sup>11</sup> <https://kraken.re/>

<sup>12</sup> Heinrich Wölfflin – *Gesammelte Werke*, <https://www.biblhertz.it/it/dept-weddigen/woelfflin>

<sup>13</sup> Le trasformazioni e sostituzioni saranno disponibili su <https://github.com/biblhertz/trans2tei>

- [4] Stefan Pletschacher and Apostolos Antonacopoulos. 2010. The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. In 2010 20th International Conference on Pattern Recognition. IEEE, 257–260. DOI: <https://doi.org/10.1109/ICPR.2010.72>
- [5] John K. G. Shearman. 2003. Raphael in early modern sources: (1483 - 1602). *Römische Forschungen der Bibliotheca Hertziana*, XXX/XXXI. Yale Univ. Press, New Haven u.a.
- [6] Phillip B. Ströbel and Simon Clematide. 2019. Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images. *Digital Humanities 2019*.
- [7] Phillip B. Ströbel, Simon Clematide, and Martin Volk. 2020. How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, 3551–3559.

# PATH, un thesaurus per le arti performative.

## Osservazioni empiriche e criticità d'uso

Michela Berti<sup>1</sup>, Manuela Grillo<sup>2</sup>

<sup>1</sup> Conservatorio Statale di Musica 'G.B. Pergolesi', Fermo, Italia – michelaberti@gmail.com

<sup>2</sup> Sapienza Università di Roma, Italia – manuela.grillo@gmail.com

### ABSTRACT<sup>1</sup>

Il presente contributo propone delle osservazioni empiriche emerse durante l'esperienza di creazione del thesaurus PATH, nato nell'ambito del progetto ERC PerformArt<sup>2</sup> come strumento di indicizzazione dei contenuti dei documenti antichi, con particolare riguardo alla sostenibilità e alle criticità emerse, non soltanto durante la fase di sviluppo, ma soprattutto durante la fase d'uso dello strumento.

### PAROLE CHIAVE

Thesaurus; organizzazione della conoscenza; indicizzazione semantica; arti performative

## 1. INTRODUZIONE

Il thesaurus PATH è stato sviluppato come strumento di organizzazione della conoscenza per il database del progetto PerformArt, dedicato alle arti dello spettacolo a Roma nei secoli XVII e XVIII, che - diretto da Anne-Madeleine Goulet (Centre d'études supérieures de la Renaissance de Tours) e coordinato da Michela Berti - ha avuto inizio il 1. Settembre 2016 ed è terminato il 31 agosto 2022.

La sfida principale del progetto è consistita nel ricostruire eventi transitori e unici, come sono gli eventi performativi, attraverso la ricerca d'archivio. La natura effimera dell'oggetto stesso della ricerca ha reso quanto mai necessario incrociare, ordinare ed elaborare automaticamente i dati, attraverso un database, la cui struttura generale è stata redatta dalla Principal Investigator fin dal momento della presentazione del progetto ed è stata progressivamente sviluppata in base alle esigenze emerse dalla ricerca.

La creazione di un thesaurus, pur prevista nelle specifiche iniziali, non era stata elaborata e dettagliata nella sua struttura e nel piano di sviluppo. Tuttavia, la grande quantità di dati raccolti ha evidenziato l'urgenza di sviluppare uno strumento dedicato: scandagliare gli archivi familiari dell'aristocrazia romana alla ricerca di documenti che contenessero traccia di eventi performativi (in conti, diari, etc.), per poi riportarli in un database dedicato, ha portato alla nascita di un corpus documentario 'creato' dall'assemblaggio dei ricercatori, in cui le singole unità documentarie sono decontestualizzate dal loro contesto. In assenza del vincolo archivistico, al contenuto informativo dei documenti vengono a mancare quegli elementi di contesto che ne qualificano il valore sia giuridico che storico, nonché la piena portata informativa.

## 2. METODOLOGIA

PATH, sviluppato a cura di Manuela Grillo, ha raggiunto le sue attuali consistenze grazie alla collaborazione stabilita da progetto nel 2018 con il Nuovo soggettario della Biblioteca Nazionale di Firenze: Nuovo Soggettario è quindi la base di sviluppo del thesaurus di PerformArt e allo stesso tempo è stato arricchito dalle proposte PerformArt, laddove il campo specifico della terminologia delle arti performative fosse carente.

Il Nuovo soggettario è un linguaggio di indicizzazione semantica adatto all'indicizzazione del contenuto di risorse di vario tipo; è stato costruito dalla Biblioteca Nazionale di Firenze in conformità ai principi IFLA (International Federation of Library Associations and Institutions) e agli standard internazionali. Come ogni linguaggio, è costituito da regole (sia per la costruzione di stringhe di soggetto che indicano il contenuto di una risorsa, sia per il controllo del vocabolario) e da un serbatoio terminologico (il thesaurus); quest'ultimo, in continua evoluzione e crescita, contiene più di 70.000 termini (con 33454 termini preferiti; ultimo aggiornamento ottobre 2022).

---

<sup>1</sup> Michela Berti ha scritto Introduzione. Manuela Grillo ha scritto Metodologia e Descrizione. Le altre parti sono state scritte a quattro mani.

<sup>2</sup> PerformArt è un programma finanziato dall'European Research Council, ospitato dal Centre National de la Recherche Scientifique, in partnership con l'École Française de Rome (Grant Agreement n. 681415).

Il Nuovo soggettario è disponibile in SKOS/RDF (Simple Knowledge Organization System, famiglia di linguaggi formali progettati per rappresentare glossari, classificazioni, tassonomie e qualsiasi tipo di vocabolario strutturato).

La scelta di partire da dati già sviluppati nel progetto Nuovo soggettario riflette l'orientamento generale della ricerca e le politiche dell'Unione Europea verso gli open data, la condivisione della conoscenza, il riutilizzo dei dati, quando questi sono validati e sviluppati secondo standard internazionali; questa scelta ha costituito un ulteriore elemento positivo quanto alla sostenibilità di tempi e costi della realizzazione dello strumento.

Il thesaurus – e l'intero database relazionale PerformArt - è stato sviluppato con il software 4D (4th Dimension), un relational database management system. Il thesaurus e il database con esso indicizzato sono consultabili all'indirizzo <https://performart.huma-num.fr/>, sulla piattaforma Huma-Num, infrastruttura digitale di livello internazionale per le scienze umane e sociali, nodo francese di DARIAH-ERIC e CLARIN: Huma-Num permette alle comunità di ricerca SSH di sviluppare, realizzare e conservare sul lungo termine dati e strumenti dei programmi di ricerca in un contesto di open science e condivisione dei dati.

### 3. DESCRIZIONE

La struttura del PATH - mutuata dal thesaurus Nuovo soggettario - comprende quattro macro-categorie, caratterizzate da tredici categorie coincidenti con altrettanti top terms, come illustrato nella tabella n. 1.

Macro-categorie	Categorie (coincidenti con i top term)
Agenti	Organismi Organizzazioni Persone e gruppi
Azioni	Attività Discipline Processi
Cose	Forme Materia Oggetti Spazio Strumenti Strutture
Tempo	Tempo

Tabella 1. Macro-struttura del thesaurus.

Al momento la struttura tesaurale del PATH è costituita da circa 2300 termini.

	Numero totale termini	Numero forme accettate (termine preferito)	Numero forme scartate (Use For)
	2286	1672	414
Termini derivanti da Nuovo Soggettario	1621	1213	211
Nuovi termini PATH (rispetto a Nuovo Soggettario)	665	459	203

Tabella 2. Consistenze PATH.

Dai dati della tabella si deduce quanto sia stato accurato il lavoro di inserimento di forme scartate per ogni nuovo termine preferito PATH: se nel Nuovo Soggettario, relativamente ai termini che sono stati importati in PATH perché necessari per indicizzare i documenti del database, abbiamo 211 forme scartate per 1213 forme preferite, durante la costruzione di PATH sono state inserite 203 nuove forme scartate per 459 nuovi termini preferiti. Le logiche di un progetto di ricerca hanno potuto garantire un lavoro capillare dal punto di vista terminologico, rispetto all'elaborazione di un thesaurus generalista: la ricchezza di forme scartate agevola l'uso dello strumento a partire dalle molte forme con cui un concetto può presentarsi nelle fonti (si veda l'esempio in figura 1 del termine preferito Fabbri che ha rinvio da sette forme scartate).



Lista		Lista gerarchica
☰		
🔍 ▲ Termine	🔍 ▼ Termine generico (BT)	🔍 ▼ Definizione (DEF)
1	<i>chiavai</i> → <i>fabbr</i>	
2	<i>chiavari</i> → <i>fabbr</i>	
3	<b>fabbr</b>	lavoratori del ferro      Artigiani che lavorano il ferro, i ferramenti
4	<i>fabbr ferrai</i> → <i>fabbr</i>	
5	<i>ferraioli</i> → <i>fabbr</i>	
6	<i>ferraiuoli</i> → <i>fabbr</i>	
7	<i>ferrari</i> → <i>fabbr</i>	
8	<i>magnani</i> → <i>fabbr</i>	
Elementi per pagina : 50 ▼		

Figura 6.

Non è comunque sufficiente un buon impianto terminologico: al di là della qualità delle gerarchie e delle relazioni tra i descrittori, è necessario stabilire una politica d'uso dei descrittori condivisa da tutta l'équipe di ricerca. Per utilizzare al meglio i descrittori del thesaurus, è stata infatti necessaria una politica di indicizzazione condivisa con tutti gli autorizzati ad inserire i descrittori, così che fossero guidati da principi di uniformità nell'assegnazione di parole chiave nei documenti oggetto delle loro ricerche; è stato quindi necessario in primis standardizzare il livello di dettaglio nell'attribuzione dei descrittori. Ad esempio, nel caso di conti contenenti i costi degli abiti di scena - in cui i costi di pizzi, fettucce e tessuti sono dettagliati - se non fosse stata stabilita una politica per l'uso di descrittori specifici, il risultato sarebbe stato molto difforme: un ricercatore avrebbe potuto inserire 'abiti di scena', un altro, più scrupoloso, 'pizzi', 'fettucce' e 'tessuti', un terzo, incerto, avrebbe scelto tutti i termini in questione. Si ritiene rilevante specificare che non esiste un livello di specificità corretto a priori (piuttosto sono le esigenze del progetto specifico a determinarlo), ma un livello di dettaglio coerente e uniforme nell'assegnazione delle parole chiave è fondamentale per un buon livello di information retrieval.

La scelta di affidare ai ricercatori l'attribuzione dei descrittori è stata dettata dal fatto che solo esperti della materia hanno la capacità di interpretare i documenti e documenti e, in questi, i loro particolari contenuti: si tratta di una enorme ricchezza per gli utenti del database, ricchezza di cui i non specialisti del settore difficilmente sarebbero riusciti a cogliere tutte le sfumature. Dopo una breve formazione sul tema indicizzazione semantica, i ricercatori sono stati affiancati da una specialista nella creazione e nell'uso di thesauri durante tutto il loro lavoro di inserimento. Questo ha comportato di fatto una concreta sostenibilità di tempi e costi di lavorazione degli item.

La principale innovazione metodologica di PATH è a livello d'uso, nell'applicazione di un linguaggio di indicizzazione per il materiale d'archivio antico.

#### 4. CONCLUSIONI

L'auspicio è che PATH possa diventare il thesaurus di riferimento per le arti sceniche durante il tardo Barocco a Roma e che possa servire per l'indicizzazione di corpus documentari diversi dai materiali di PerformArt, considerato il suo vasto potenziale come serbatoio di vocabolario specifico per le arti dello spettacolo.

Inoltre per sviluppare il PATH è stato necessario studiare le esperienze già condotte o in corso, relative alla semantica musicale: è così emersa la scarsa cultura della standardizzazione in musicologia, la cui diretta conseguenza è la scarsa condivisione dei dati e la modestissima presenza di dati musicali nel web semantico, quindi l'ulteriore auspicio è che l'esperienza del PATH sia il primo passo verso un dialogo tra progetti realizzati in realtà diverse a livello europeo e internazionale.

#### BIBLIOGRAFIA

- [1] Bardiot, Clarisse, Arts de la scène et humanités numériques: des traces aux données. Iste, 2021.
- [2] Berti, Michela e Manuela Grillo. «Digital tools for the study of historical performing arts: the PerformArt database and thesaurus». UD Umanistica Digitale 10 (2021): 443-450. DOI: <https://doi.org/10.6092/issn.2532-8816/12624>.
- [3] Berti, Michela e Manuela Grillo, «From scientific research to a difficult definition, and back again: Magnificenza as a paradigm in the development of a thesaurus». In Noble Magnificence : Culture of the Performing Arts in Rome, 1644-1740, edited by Anne-Madeleine Goulet e Michela Berti, Brepols, (in uscita, collezione «Epitome musical»).
- [4] Biagetti, Maria Teresa, Ontologie bibliografiche. Bulzoni, 2022.

- [5] Blough, Kathy, and Karla Jurgemeyer. «Library of Congress Genre/Form Terms and Library of Congress Medium of Performance Thesaurus for Music». *Music Reference Services Quarterly* 2 (2015): 122–125.
- [6] Bonora, Paolo e Angelo Pompilio. «RePIM in LOD: semantic technologies to manage, preserve, and disseminate knowledge about Italian secular music and lyric poetry from the 16th-17th centuries». *Umanistica Digitale* 14 (2022). DOI: <http://doi.org/10.6092/issn.2532-8816/15568>
- [7] Broughton, Vanda e Aida Slavic. «Building a faceted classification for the humanities: principles and procedures». *Journal of documentation* 5 (2007): 727-754.
- [8] Goulet, Anne-Madeleine. «Jeux d'échelle. De l'intérêt d'un financement ERC pour la recherche en SHS». *La lettre de l'InSHS* 7 (2019): 20-22.
- [9] Goulet, Anne-Madeleine. «Un outil pour étudier les spectacles de l'Ancien Régime. La base de données PerformArt». *Revue d'Historiographie du Théâtre* 5 (2020). <https://sht.asso.fr/un-outil-pour-etudier-les-spectacles-de-lancien-regime-la-base-de-donnees-performart-2/>
- [10] Grillo, Manuela. «Costruzione del thesaurus, analisi a faccette, categorie». *Biblioteche oggi* 6 (2007): 97-103.
- [11] Grillo, Manuela. *Indicizzazione semantica di bandi, manifesti e fogli volanti*. Documenta, 2015.
- [12] Grillo, Manuela. «La costruzione di un thesaurus per la performance». In *Spectacles et performances artistiques à Rome (1644-1740). Une analyse historique à partir des archives familiales, sous la direction de Anne-Madeleine Goulet, Élodie Oriol, José Maria Dominguez*. École française de Rome, 2021. DOI: 10.4000/books.efr.16344
- [13] Guerrini, Mauro e Tiziana Possemato. *Linked data per archivi, biblioteche, musei. Perché l'informazione sia del web e non solo sul web*. Editrice Bibliografica, 2015.
- [14] Hemmasi, Harriette. «The Music Thesaurus: Function and Foundations». *Notes: Quarterly Journal of the Music Library Association* 3 (1994): 875–882.
- [15] Lucarelli, Anna. «Biblioteche digitali e specialisti del mondo antico. Collaborazioni nell'ambito dell'indicizzazione dell'indicizzazione e della ricerca semantica». In *Strumenti digitali e collaborativi per le Scienze dell'antichità*, a cura di Paolo Mastandrea. Edizioni Ca' Foscari, 2017, 219-229.
- [16] Lucarelli, Anna. «Un'opera è un'opera: dialogo, non troppo immaginario, sull'indicizzazione dell'antico». In *Viaggi a bordo di una parola. Scritti sull'indicizzazione semantica in onore di Alberto Cheti*, a cura di Anna Lucarelli, Alberto Petrucciani, Elisabetta Viti. AIB, 2019, 141-161.
- [17] Marangi, Valentina Mirella. «Verso un thesaurus in materia di Legislazione sui beni e attività musicali: proposte di integrazione al TEsauro Senato per l'Organizzazione dei documenti parlamentari». *Fonti Musicali Italiane: Periodico di Ricerca Musicologica* 6 (2001): 145–159.
- [18] Piéjus, Anne e Nathalie Berton-Blivet, Thomas Bottini, «De la mise en récit de l'histoire artistique au système d'information numérique partagé. Le programme Mercure galant». *Revue d'Historiographie du Théâtre* 5 (2020). <https://sht.asso.fr/de-la-mise-en-recit-de-lhistoire-artistique-au-systeme-dinformation-numerique-partage-le-programme-mercure-galant/>
- [19] Schneider, Klaus. «Über Den Musik-Thesaurus Und Die Inhaltliche Erschließung von Musik». *Forum Musikbibliothek: Beiträge Und Informationen Aus Der Musikbibliothekarischen Praxis* 2 (1996): 100–108.
- [20] Spilker, John. «Toward an International Music Thesaurus». *Fontes Artis Musicae* 1 (2005): 29–44.

# I dati della ricerca: gestione, diffusione e preservazione nei Next generation repositories

Ilaria Belvedere<sup>1</sup>, Simona Turbanti<sup>2</sup>

<sup>1</sup> Biblioteca Tiziano Terzani, Campi Bisenzio (Firenze), Italia – [ilaria.belvedere9@gmail.com](mailto:ilaria.belvedere9@gmail.com)

<sup>2</sup> Università di Pisa, Italia – [sturbanti@gmail.com](mailto:sturbanti@gmail.com)

## ABSTRACT

Data la crescente varietà dei dati prodotti dalle comunità di ricerca, la gestione di queste informazioni “granulari” non può prescindere dall’implementazione di piattaforme in grado di garantire la loro conservazione, visibilità e riuso, orientate verso il modello dei Next generation repositories. Tra le soluzioni disponibili, il software open source Dataverse, pensato specificatamente per i dati della ricerca, permette di strutturare i contenuti in archivi virtuali all’interno dei quali i dati possono essere gestiti a tutto tondo, grazie alla disponibilità di metadati descrittivi generali e specifici, di un sistema codificato di citazione e di formati diversi di visualizzazione ed esportazione. Un’analisi condotta presso l’Università di Pisa, attraverso una versione demo del software in locale denominata DataversePI, ha mostrato le potenzialità dello strumento e i vantaggi che porterebbe all’istituzione in termini di conservazione, preservazione, accesso e valorizzazione delle informazioni create durante tutte le fasi della ricerca. I Next generation repositories delle diverse istituzioni possono formare, inoltre, una rete distribuita in cui i diversi archivi interagiscono tra loro, pur mantenendo ognuno le proprie specificità.

## PAROLE CHIAVE

Dati della ricerca; data management plans; accesso aperto; archivi istituzionali; Dataverse

## 1. INTRODUZIONE<sup>1</sup>

Gli enti di ricerca si trovano, ormai da decenni, a dover mettere in atto strategie per la gestione dei dati della ricerca, ovvero di quei prodotti informativi generati «durante tutto il ciclo di ricerca [che] includono qualsiasi informazione in formato binario digitale [...] creata, immagazzinata, acceduta e restituita con l’uso della tecnologia dei computer» ([9]), comprendenti pertanto sia le pubblicazioni che i materiali non pubblicati.

Come sottolineato nelle *Linee guida sull’interoperabilità tecnica delle pubbliche amministrazioni*<sup>2</sup>, all’interno del Piano triennale per l’informatica nella Pubblica amministrazione, «i dati pubblici sono un bene comune e una risorsa del Paese in grado di produrre valore migliorando i servizi, creandone di innovativi e contribuendo a creare nuovi modelli di business, competenze e posti di lavoro»<sup>3</sup>. Una gestione più efficace dei dati pubblici, la creazione di servizi focalizzati sui cittadini agevolati dall’interoperabilità tra enti, il supporto all’attività imprenditoriale e alla ricerca scientifica rappresentano i principali obiettivi del Dipartimento per la trasformazione digitale del Governo italiano.

La pianificazione delle attività di *Research data management* (RDM) è necessaria anche in vista delle direttive internazionali in materia di gestione dei dati. Il Programma quadro dell’Unione europea per la ricerca e l’innovazione Horizon 2020<sup>4</sup> – in vigore per gli anni 2014-2020 – attraverso l’Open research data pilot<sup>5</sup> ha previsto l’obbligo dell’accesso aperto dove possibile, secondo il principio «as open as possible, as close as necessary» per i risultati della ricerca finanziati nell’ambito del programma pensati per la pubblicazione in riviste *peer-reviewed*: si prescrive, quindi, di depositare le pubblicazioni in archivi di dati della ricerca e di redigere i relativi *Data management plans* (DMPs).

Il documento FAIR *guiding principles* ([15]), elaborato nel 2016, presenta quattro importanti principi, che non rappresentano standard, ma raccomandazioni da seguire per una corretta gestione dei dati della ricerca: quest’ultimi dovrebbero essere resi rintracciabili, accessibili, interoperabili e riutilizzabili (*findable, accessible, interoperable,*

<sup>1</sup> Questo contributo prende spunto dalla tesi magistrale in Informatica umanistica di Ilaria Belvedere dal titolo I Research Data Repositories nella gestione dei dati della ricerca: ipotesi di implementazione di un repository Dataverse per l’Università di Pisa (relatore prof.ssa Simona Turbanti, correlatore prof. Paolo Milazzo), discussa presso l’Università di Pisa il 29 settembre 2022.

<sup>2</sup> Cfr. Italia, Presidenza del Consiglio dei ministri, Dipartimento per la trasformazione digitale, *Linee guida sull’interoperabilità tecnica delle pubbliche amministrazioni*. <https://docs.italia.it/italia/piano-triennale-ict/lg-modellointeroperabilita-docs/it/bozza/index.html>

<sup>3</sup> Cfr. <https://innovazione.gov.it/progetti/dati-e-interoperabilita/>

<sup>4</sup> Cfr. [https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020\\_en](https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en)

<sup>5</sup> Cfr. <https://www.openaire.eu/what-is-the-open-research-data-pilot>

reusable). I principi FAIR non potrebbero articolarsi senza la presenza di una policy sui dati della ricerca stilata dall'istituzione e senza il *Data management plan*, documento necessario al momento della presentazione di un progetto di ricerca.

Secondo quanto definito proprio all'interno del documento *Guidelines on FAIR data management in Horizon 2020* ([7]), i DMPs sono le descrizioni del ciclo di gestione dei dati della ricerca in uno specifico progetto di ricerca e costituiscono un elemento cardine per una buona gestione dei dati. Horizon Europe<sup>6</sup>, che fa seguito a Horizon 2020 per gli anni 2021-2027, ribadisce quanto previsto in precedenza puntando all'apertura di tutti i dati della ricerca.

## 2. GLI ARCHIVI ISTITUZIONALI DELLA RICERCA

Nonostante la letteratura presenti da oltre un ventennio gli innumerevoli apporti positivi degli archivi istituzionali della ricerca (Institutional repositories, IR) ([10]), spesso gli studiosi percepiscono il deposito dei propri "prodotti" - metadati e full text - all'interno di questi strumenti solo come un ulteriore obbligo cui assolvere ai fini degli esercizi valutativi locali e nazionali delle università/enti di ricerca senza cogliere le opportunità, anche a livello personale, che ne derivano, in termini di visibilità del proprio lavoro. L'agevole reperibilità in rete dei dati descrittivi e l'accesso al testo pieno del contributo, se possibile e nella forma eventualmente consentita dalle politiche degli editori (versione editoriale, preprint, postprint), consente, infatti, una maggiore circolazione dei risultati e, per giovani studiosi, la possibilità di attivare legami con ricercatori di altre realtà geografiche e di "fare rete".

La crescente produzione di dati di diverse tipologie ha dettato, inoltre, la necessità di pensare a modelli nuovi di archivi che tengano conto della complessità di queste risorse. Nel 2017 è stato pubblicato il report ([11]) dei risultati ottenuti dal gruppo di lavoro Next Generation Repository Working Group della Confederation of Open Access Repositories (COAR)<sup>7</sup>, fondamento di un'infrastruttura distribuita e globalmente connessa a beneficio di ricerca, innovazione e collaborazione della comunità accademica. Uno degli elementi più importanti di questo programma consiste nel prevedere che gli IR conservino materiali variegati, così da permettere una condivisione di tipologie diverse di prodotti della ricerca e un loro riconoscimento.

I Next generation repository considerano sia i metadati sia i prodotti della ricerca come risorse web da identificare tramite URL, avvalendosi quindi dei vocabolari controllati propri del web semantico con un approccio pragmatico. Il report presenta sei principi guida e il modello concettuale alla base della costituzione dei NGR; i principi definiscono come essenziale la presenza di un sistema distribuito, cioè di una rete di repositories che eviti il controllo monopolistico della singola istituzione e, al contempo, garantisca la tutela delle specificità di ogni contesto, puntando all'accessibilità delle risorse ([10, p. 8]).

Anche a livello europeo si punta l'accento sull'interazione tra istituzioni; in particolare, per quanto riguarda il settore della ricerca, l'interoperabilità tra le infrastrutture è il focus del documento pubblicato dalla European Open Science Cloud (EOSC)<sup>8</sup>, EOSC interoperability framework report ([6]), frutto del lavoro di EOSC FAIR Working Group<sup>9</sup> e EOSC Architecture Working Group<sup>10</sup>: si sottolinea l'importanza dell'interoperabilità per la gestione dei dati della ricerca, distinguendo tra interoperabilità tecnica, semantica, organizzativa e legale. Attraverso uno studio teorico e pratico, condotto anche tramite interviste ad alcune comunità di ricerca appartenenti a EOSC, sono state individuate problematiche, necessità e raccomandazioni da seguire per ognuno dei quattro tipi di interoperabilità. Nel programma di lavoro di Horizon Europe per l'anno 2023-2024 vengono ribadite queste considerazioni.

## 3. I NEXT GENERATION REPOSITORIES: IL SOFTWARE OPEN SOURCE DATAVERSE

Tenendo presente il modello dei Next generation repositories, un ente di ricerca che voglia implementare un archivio pensato specificatamente per la gestione dei dati della ricerca deve considerare diversi fattori. Durante la fase di pianificazione, oltre all'architettura della piattaforma e le sue caratteristiche ([11]), è importante vagliare anche le funzionalità di gestione, preservazione e accessibilità dei contenuti e l'integrazione con servizi esterni. A questo proposito, il recente documento *The TRUST principles for digital repositories* ([3]) ha presentato alcuni principi guida per la configurazione dei repository per una gestione dei dati FAIR: i primi due riguardano la trasparenza in merito ai servizi offerti dal repository e al deposito dei dati, e la responsabilità collegate, il terzo è relativo ai bisogni dell'utenza, mentre il quarto e il quinto sono incentrati sulla preservazione dei dati nel lungo periodo e sugli aspetti tecnologici.

---

<sup>6</sup> Cfr. <https://horizoneurope.apre.it/>

<sup>7</sup> Cfr. <https://www.coar-repositories.org/>

<sup>8</sup> Cfr. <https://eosc-portal.eu/>

<sup>9</sup> Cfr. <https://eoscsecretariat.eu/working-groups/fair-working-group>

<sup>10</sup> Cfr. <https://eoscsecretariat.eu/working-groups/architecture-working-group>

In riferimento alla scelta del software, tra le opzioni open source disponibili, si annoverano Eprints<sup>11</sup> e Dspace<sup>12</sup> da una parte, validi anche se pensati prevalentemente per la gestione solo delle pubblicazioni, e soluzioni come InvenioRDM<sup>13</sup>, Haplo<sup>14</sup> e Dataverse<sup>15</sup>. Dataverse, in particolare, è stato scelto da diverse istituzioni internazionali e da alcune realtà italiane come l'Università degli studi di Milano e l'Istituto italiano di tecnologia (IIT).

Dataverse, nato nel 2006 dalla collaborazione tra l'Institute for Quantitative Social Science (IQSS) dell'Università di Harvard e altre istituzioni e giunto attualmente alla versione 5.13, si presenta come uno strumento per la gestione dei dati della ricerca che consente un uso in autonomia ai fruitori e senza una specifica connotazione disciplinare; la sua architettura si basa su Java Enterprise Edition (JEE)<sup>16</sup>. Grazie al suo design modulare il software consente di contribuire al codice, di personalizzare le installazioni secondo le necessità delle istituzioni e di avere un nucleo funzionale ([4]): dalla versione 4.8.6 è stato introdotto un *external tool framework* per permettere l'integrazione con programmi esterni senza intervenire nel *core code*, come in passato ([5]).

L'intento del progetto Dataverse consiste nell'offrire ai ricercatori uno spazio dove poter caricare i contenuti della ricerca che consenta di assicurare, da una parte, visibilità, sicurezza e garanzie di preservazione ai dati, dall'altra personalizzazione e autonomia di gestione per i professionisti della ricerca.

L'organizzazione dei contenuti, suddivisi in archivi virtuali chiamati *Dataverse collections* che, a loro volta, possono contenere altre collezioni o datasets, è intuitiva e facilmente gestibile sia tramite interfaccia che utilizzando le API. In Dataverse i dati della ricerca, conservati sul server istituzionale, sono visibili grazie a un sistema formale di citazione - a livello di file e dataset, in fase di pianificazione per le collezioni -, interoperabili (protocollo OAI-PMH<sup>17</sup>), ricercabili e accessibili grazie alle *features* relative ai metadati ([2]) ([14]), analizzabili attraverso strumenti specifici e facilmente integrabili all'interno di pagine web esterne. Per quel che concerne in particolare i metadati, Dataverse ne supporta l'inserimento sia a livello di collezioni, sia a livello dataset (fig. 1) e file.



Figura 7. Le funzionalità relative ai metadati livello dataset.

Il software prevede l'utilizzo di standard internazionali, come Dublin core (DC)<sup>18</sup> e Documentation data initiative (DDI)<sup>19</sup>, e si avvale della collaborazione con DataCite<sup>20</sup> e Google dataset search<sup>21</sup>. Inoltre, come osservato da Bantin ([1]), la pubblicazione dei dati e dei metadati avviene separatamente, garantendo la sicurezza dei dati. Altre importanti strumenti offerte da Dataverse riguardano:

- la gestione dei ruoli degli utenti in relazione alle collezioni e ai datasets;

<sup>11</sup> Cfr. [https://wiki.eprints.org/w/Main\\_Page](https://wiki.eprints.org/w/Main_Page)

<sup>12</sup> Cfr. <https://dspace.lyrasis.org/>

<sup>13</sup> Cfr. <https://inveniordm.web.cern.ch/>

<sup>14</sup> Cfr. <https://support.haplo.com/repository>

<sup>15</sup> Cfr. <https://dataverse.org/about>

<sup>16</sup> Cfr. <https://www.oracle.com/java/technologies/java-ee-glance.html>

<sup>17</sup> Cfr. [https://it.wikipedia.org/wiki/Open\\_Archives\\_Initiative\\_Protocol\\_for\\_Metadata\\_Harvesting](https://it.wikipedia.org/wiki/Open_Archives_Initiative_Protocol_for_Metadata_Harvesting)

<sup>18</sup> Cfr. <https://www.dublincore.org/>

<sup>19</sup> Cfr. <https://ddialliance.org/>

<sup>20</sup> Cfr. <https://datacite.org/>

<sup>21</sup> Cfr. <https://datasetsearch.research.google.com/>

- la scelta di licenze d'uso o la creazione di propri termini d'uso personalizzati,
- il supporto per i Big data;
- la preservazione dei dati;
- l'integrazione con servizi esterni, per esempio con piattaforme per l'analisi dei dati.

#### 4. I VANTAGGI DI DATAVERSE

Per avere un riscontro concreto sull'implementazione di un repository Dataverse, sono stati contattati 17 rappresentanti di istituzioni che hanno adottato il software e analizzato le fonti dirette di una istituzione, per un totale di 18 casi studio. L'obiettivo è stato individuare le motivazioni dietro alla loro scelta, riassunte in un istogramma e relativa tabella (fig. 2).

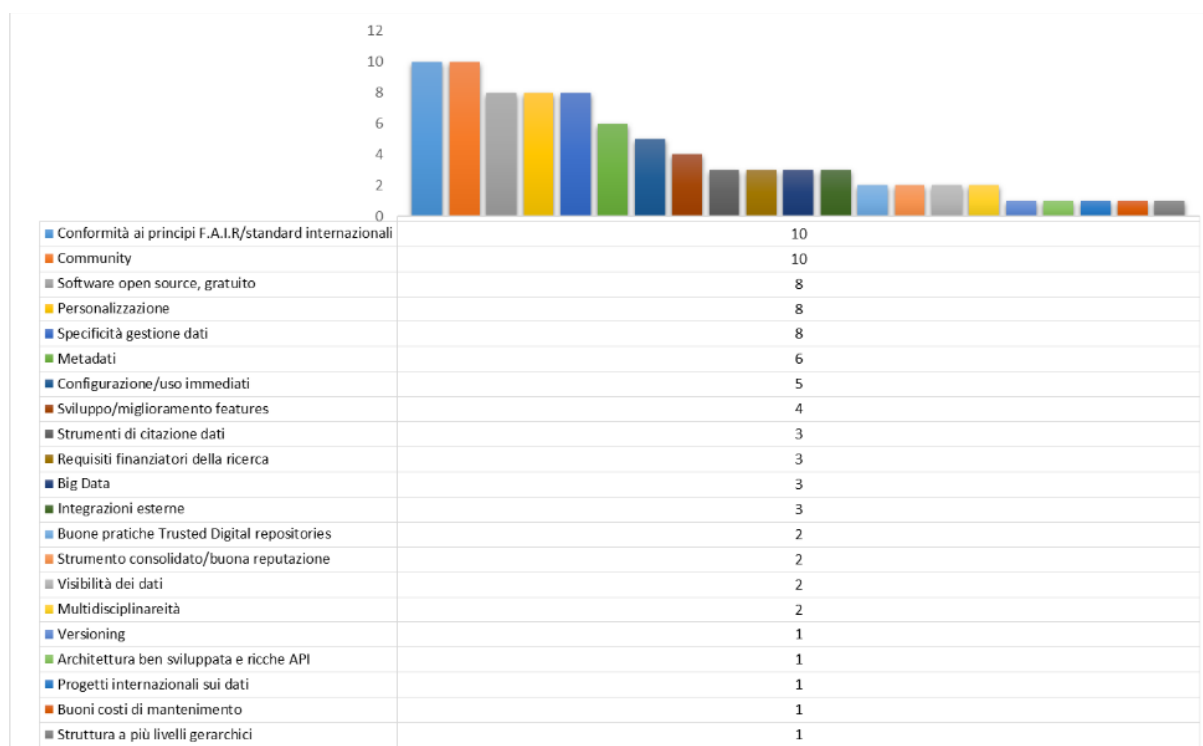


Figura 8. Feedback istituzioni.

Dalle interviste e dalle fonti è emerso che Dataverse potrebbe essere una buona soluzione per diverse ragioni. Le motivazioni più citate dalle istituzioni contattate sono, a pari merito, il fatto che Dataverse permetta una gestione dei dati della ricerca conforme ai principi FAIR e agli standard internazionali, il costante impegno della comunità di sviluppatori e contributori, oltre al supporto e disponibilità costanti verso gli utilizzatori del software. Gli altri tre vantaggi più menzionati riguardano la gratuità e la piena disposizione del codice, l'alto grado di personalizzazione in riferimento sia all'interfaccia, sia alla parte contenutistica e alla gestione dei ruoli e dei permessi, e la specificità, ovvero l'essere stato progettato in funzione dei *research data*. Un'altra importante ragione riportata da 6 istituzioni è l'attenzione ai metadati, la possibilità di creare blocchi di metadati personalizzabili, di utilizzare standard specifici e di effettuare l'*harvesting*. Nonostante la sua installazione e personalizzazione comportino certamente un costo di gestione, soprattutto in termini di risorse umane, le istituzioni coinvolte affermano che la configurazione e l'utilizzo del software sono immediati. Tra gli altri vantaggi segnalati, l'impegno della community nel migliorare le funzionalità già esistenti e nell'implementarne *ex-novo* e la presenza di strumenti di citazione specifici per i dati; peraltro, quest'ultimo elemento si inserisce nell'ambito della gestione dei dati e si collega alla già citata "visibilità" dei dati della ricerca. La comunità di Dataverse realizza collaborazioni in merito alla gestione dei dati della ricerca con fornitori di servizi analoghi, perseguendo l'interoperabilità, come chiaramente esplicitato nella Roadmap. In considerazione di ciò, il software può risultare un valido strumento per la realizzazione non soltanto di singoli repository di dati, ma anche di articolate infrastrutture di ricerca.



## 5. UN PROTOTIPO PER L'UNIVERSITÀ DI PISA: DATAVERSEPI

Al fine di testare direttamente le funzionalità di Dataverse è stato condotto uno studio, focalizzato sull'Università di Pisa, implementando una versione *demo* del software in locale, denominata DataversePI. Dopo l'installazione su una macchina virtuale di Ubuntu 18.04.6, è stato scaricato il pacchetto Archive in a box<sup>22</sup>, che consente una modalità di installazione immediata di Dataverse nella sua versione 5.11.1 avvalendosi del modulo Dataverse Docker, ideato da Data Archiving and Networked Services (DANS)<sup>23</sup>, Royal Netherlands Academy of Arts and Sciences (KNAW)<sup>24</sup> e Dutch Research Council (NWO)<sup>25</sup>. Archive in a box consente di servirsi delle funzioni più importanti di Dataverse e di avvalersi del supporto di vocabolari controllati esterni, di identificatori e DOI<sup>26</sup>, della personalizzazione di metadati e di altri strumenti offerti dal software nella sua versione classica.

La strutturazione dei contenuti ha previsto, a un primo livello, la *root collection*, pensata per contenere tutte le altre collezioni, a un secondo livello, le collezioni dei dipartimenti dell'Università di Pisa, le quali, a loro volta, ospitano le *collections* del livello successivo, corrispondenti agli archivi dei progetti di ricerca, dei gruppi di ricerca, dei laboratori di ricerca e dei ricercatori affiliati ai dipartimenti. Infine, a un quarto livello, sono presenti i dataset delle collezioni. All'interno del caso studio è stata creata la collezione del Dipartimento di filologia, letteratura e linguistica.

Per la gestione delle collezioni e dei loro contenuti sono state testate sia l'interfaccia sia la NATIVE API, un tipo di API di facile utilizzo anche per i non sviluppatori. La pagina della *root collection*, corrispondente alla homepage di DataversePI, presenta una maschera di ricerca *Google-like* e una funzione di ricerca avanzata. Sono disponibili, inoltre, filtri di visualizzazione e faccette di navigazione impostati sulla base dei metadati di citazione; oltre ai tre campi assegnati di default – Dataverse category (che indica la tipologia di Dataverse, per esempio Department), Publication year e Publication status – sono stati selezionati quelli ritenuti più significativi, ossia Subject, Keyword term, Type of data, Language, Grant information Grant agency, Grant information, Grant number e Deposit date.

Oltre ai metadati citazionali, le informazioni scelte per il Dipartimento di filologia, letteratura e linguistica sono Geospatial metadata, Social science and humanities Metadata e Journal metadata, in linea con le discipline studiate dai membri del Dipartimento; sono state, inoltre, mantenute le faccette di navigazione della *root collection*.

Il Dataverse FILELI ospita al suo interno altri sei Dataverse:

- 3 Dataverse di tipo “ricercatore”;
- 2 Dataverse di tipo “laboratori di ricerca”;
- 1 Dataverse di tipo “gruppo di ricerca”

A loro volta i sei Dataverse contengono ciascuno almeno un dataset; per esempio, ognuno dei tre Dataverse dei ricercatori include un dataset contenente diversi tipi di file, in particolare file tabulari e file di testo.

Nella pagina di visualizzazione del dataset, oltre alla citazione e ai suoi metadati, è possibile esplorare (ed editare, nel caso si abbia un ruolo che lo consente) i file, i metadati dei file, i tag relativi alla loro tipologia, i loro path, i termini di utilizzo e la versione, dato che il dataset mantiene la storia delle sue modifiche. Tramite il tasto Access dataset o utilizzando la DATA ACCESS API il dataset è scaricabile in formato zip.

Al livello dei file, il software fornisce automaticamente una citazione che può essere esportata in vari formati, così come anche i metadati relativi sono esportabili nei formati più diffusi<sup>27</sup>. Sia per i dataset che per i file è possibile gestire i termini d'uso e l'accesso ai dati.

Mediante la funzionalità Guestbooks il Download users FILELI chiede obbligatoriamente nome, cognome, mail e istituzione a cui la persona è affiliata, mentre lascia opzionale il campo Position; prevede, inoltre, una domanda a risposta multipla, obbligatoria, sugli scopi di utilizzo del materiale che si andrà a scaricare.

Il test diretto delle principali funzionalità di Dataverse ha avuto il principale scopo di sottoporre l'idea dell'adozione del software per un repository dei dati della ricerca all'Università di Pisa, in particolare agli organi deputati al supporto della ricerca, ovvero il Sistema bibliotecario di ateneo<sup>28</sup> e la Direzione Servizi per la ricerca e il trasferimento tecnologico<sup>29</sup>. Il repository Dataverse Unipi sarebbe un archivio istituzionale dei dati della ricerca che, insieme alle piattaforme universitarie

<sup>22</sup> Cfr. <https://github.com/IOSS/dataverse-docker>

<sup>23</sup> Cfr. <https://dans.knaw.nl/en/>

<sup>24</sup> Cfr. <https://www.knaw.nl/en>

<sup>25</sup> Cfr. <https://www.nwo.nl/>

<sup>26</sup> Cfr. <https://www.doi.org/>

<sup>27</sup> Cfr. <https://guides.dataverse.org/en/latest/user/dataset-management.html>

<sup>28</sup> Cfr. <http://www.sba.unipi.it/>

<sup>29</sup> Cfr. <https://www.unipi.it/index.php/amministrazione/itemlist/category/2018-direzione-sprtto>

pisane già esistenti, ovvero all'Archivio della ricerca (ARPI)<sup>30</sup> e all'Archivio digitale delle tesi discusse (ETD)<sup>31</sup>, permetterebbe alla comunità pisana di conservare, preservare e condividere i prodotti derivanti dalle varie fasi del lavoro di ricerca che andrebbero ad aggiungersi alle pubblicazioni depositate in ARPI.

## 6. UNA RETE DI DATA STATIONS

All'interno delle strategie e disposizioni internazionali in merito all'Open science, negli ultimi anni l'attenzione è stata rivolta alle cosiddette *Open Science Infrastructures* (OSInfras) ([8]). Le raccomandazioni UNESCO in merito all'Open science ([13]) individuano queste ultime in tutte le infrastrutture di ricerca condivise, fisiche e virtuali, che sono essenziali per la Scienza aperta, comprendendo nel novero anche gli archivi dei dati della ricerca. Un esempio di infrastruttura di ricerca in ambito europeo è OPERAS<sup>32</sup>, dedicata al supporto della comunicazione accademica aperta nell'ambito delle scienze umane e sociali. Come espressamente indicato nella descrizione del progetto<sup>33</sup>, OPERAS è focalizzata sulla comunicazione e pubblicazione scientifiche nelle suddette discipline e persegue gli obiettivi dell'Open science.

Il centinaio di archivi digitali di dati della ricerca<sup>34</sup> che, a livello internazionale, utilizzano Dataverse, possono essere strutturati in modo diverso: in alcuni casi, il repository è l'espressione unica di una singola istituzione, in altri è progettato come un repository federato con un'unica interfaccia contenente le collezioni dei diversi enti di ricerca - quindi di più istituzioni - sotto forma di Dataverse *collections*. Un'altra soluzione è creare una rete distribuita di repository Dataverse costituendo vere e proprie infrastrutture di ricerca. Per esempio, il centro nazionale di ricerca olandese DANS, che mira alla riusabilità dei dati della ricerca, comprende nella sua nuova infrastruttura:

- le cosiddette *Data stations*, ciascuna riferita a uno specifico ambito disciplinare, ovvero istanze di Dataverse in cui i ricercatori possano gestire i loro dati sotto la supervisione di un Data station manager;
- DataverseNL<sup>35</sup>, una piattaforma aperta di archiviazione dei dati della ricerca che a sua volta contiene Dataverse *collections* e datasets di varie comunità di ricerca.

Similmente, nell'ambito del progetto Social sciences & humanities open cloud (SSHOC)<sup>36</sup>, finanziato da Horizon 2020, è previsto l'utilizzo del software Dataverse per la realizzazione di una rete di *repositories* per le organizzazioni di ricerca dell'area delle scienze sociali e umanistiche di EOSC, come il Consortium of European Social Science Data Archives (CESSDA)<sup>37</sup> e il Common Language Resources and Technology Infrastructure (CLARIN)<sup>38</sup>. DANS partecipa come partner al progetto insieme a CESSDA e ha un ruolo di primo piano nel fornire l'infrastruttura necessaria. Le *Data stations* possono essere implementate sul cloud di EOSC oppure come singole installazioni utilizzando il già citato pacchetto Archive in a box. Il progetto usa il proxy Traefik, che è *cloud-based* e permette un collegamento con le altre realtà Dataverse.

Per raggiungere una maggiore interoperabilità semantica, in SSHOC si sta lavorando a mappature tra i diversi schemi di metadati, al fine di collegare i metadati alle ontologie, rinforzare il supporto all'utilizzo dei vocabolari controllati e rendere ancora più *semantic-friendly* gli strumenti a disposizione. Come sottolineato da Slava Tykhonov, Marion Wittenberg e Wilko Steinhoff, la realizzazione di un sistema distribuito Dataverse assicura il dialogo tra le *Data stations* permettendo a ciascuna, allo stesso tempo, di mantenere le proprie caratteristiche specifiche in base alle esigenze del contesto istituzionale ([12]).

## 7. SCENARI E OPPORTUNITÀ FUTURI

Nel complesso panorama della gestione dei dati della ricerca, rappresentati da oggetti diversi quali parole, numeri, formule, grafici, tabelle, riferimenti bibliografici, software, audio, ecc., diventa sempre più importante per le istituzioni poter contare su strumenti adeguati alla gestione dei dati secondo i principi FAIR di reperibilità, accessibilità, interoperabilità e riutilizzo. Si sostiene che i repository, a causa dell'utilizzo di tecnologie e protocolli progettati oltre due decenni fa, prima dell'esplosione del web, degli strumenti social, del web semantico e dell'uso predominante dei dispositivi mobili per accedere ai contenuti in rete, non abbiano sviluppato del tutto il loro potenziale ([11]); come ricordato in precedenza, tra i

---

<sup>30</sup> Cfr. <https://arpi.unipi.it>

<sup>31</sup> Cfr. <https://etd.adm.unipi.it>

<sup>32</sup> Cfr. <https://operas-eu.org/>

<sup>33</sup> Cfr. <https://operas-eu.org/about/operas-in-a-nutshell/>

<sup>34</sup> Cfr. <https://dataverse.org/installations>

<sup>35</sup> Cfr. <https://dataverse.nl/>

<sup>36</sup> Cfr. <https://sshopencloud.eu/project>

<sup>37</sup> Cfr. <https://www.cessda.eu/>

<sup>38</sup> Cfr. <https://www.clarin.eu/>



motivi di questa mancata occasione si annovera anche l'atteggiamento spesso passivo degli studiosi utilizzatori che non ne hanno colto le potenzialità, forse anche in quanto non sensibilizzati adeguatamente da parte di soggetti e organi deputati. A questo punto, lo sviluppo di piattaforme istituzionali che si avvalgono di software specifici in continuo aggiornamento e gestiti da ampie comunità di utenti potrebbe gettare le basi per la nascita di uno "spazio" in grado di aggregare i dati prodotti dalle comunità scientifiche a livello nazionale o, addirittura, internazionale, così da aprire nuovi scenari per un sistema sempre più "ricerca-centrico", in linea con la visione di COAR: «Our vision is to position repositories as the foundation for a distributed, globally networked infrastructure for scholarly communication, on top of which layers of value added services will be deployed, thereby transforming the system, making it more research-centric, open to and supportive of innovation, while also collectively managed by the scholarly community»<sup>39</sup>.

## BIBLIOGRAFIA

- [1] Bantin, Philip C., a c. di. *Building Trustworthy Digital Repositories: Theory and Implementation*. Rowman & Littlefield, 2016.
- [2] Crosas, Mercè. «Fair principles and beyond: implementation in Dataverse». *Septentrio conference series*, n. 2 (20 marzo 2020). <https://doi.org/10.7557/5.5334>
- [3] Dawey, Lin, Jonathan Crabtree, Ingrid Dillo et al. «The TRUST principles for digital repositories». *Scientific data*, 7 (2020). <https://doi.org/10.1038/s41597-020-0486-7>
- [4] Durand, Gustavo. «Dataverse's Approach to Technical Community Engagement». *Septentrio conference series*, n. 2 (20 marzo 2020). doi: <https://doi.org/10.7557/5.5424>
- [5] Durand, Gustavo. «Integrations and External Tools». Presentato al *Dataverse Community Meeting 2021*, giugno 2021.
- [6] European Commission, Directorate-General for Research and Innovation, Corcho, Eriksson, Kurowski et al. *EOSC interoperability framework : report from the EOSC Executive Board Working Groups FAIR and Architecture*. 2021. <https://doi.org/10.2777/620649>
- [7] European Commission, Directorate-General for Research & Innovation. «H2020 Programme: Guidelines on FAIR data management in Horizon 2020 Version 3.0.» Report. European Commission, Directorate-General for Research & Innovation, 2016. <https://repository.oceanbestpractices.org/handle/11329/1259>
- [8] Grieco, Giuseppe, Ivan Heibi, Arcangelo Massari, Arianna Moretti, e Silvio Peroni. «Enabling portability and reusability of open science infrastructures.» In: *Linking theory and Practice of digital libraries*. TPDL 2022. Lecture notes in computer science, vol 13541. Springer, Cham. [https://doi.org/10.1007/978-3-031-16802-4\\_36](https://doi.org/10.1007/978-3-031-16802-4_36)
- [9] Higgins, Sarah. «The lifecycle of data management». In *Managing research data*, a cura di Graham Pryor, 17–46. Facet, 2012. <https://doi.org/10.29085/9781856048910.003>
- [10] Jain, Priti. «New trends and future applications/directions of institutional repositories in academic institutions». *Library review* 60 (2011): 125–41.
- [11] Rodrigues, Eloy, et al. «Next generation repositories: behaviours and technical recommendations of the COAR Next Generation Repositories Working Group», 28 novembre 2017. <https://doi.org/10.5281/zenodo.1215014>
- [12] Tykhonov, Vyacheslav, Marion Wittenberg e Wilko Steinhoff. «Archive in a box for Dataverse: building distributed data network with shared services». *Dataverse Community Meeting 2022*, 14 giugno 2022, Software session 1, <https://projects.iq.harvard.edu/dcm2022/agenda>, [https://www.youtube.com/watch?v=-Re\\_elg67e8&list=PLLoKvRqQVbtKJFde-bynhnXdePwI7jpS&index=2&t=4378s](https://www.youtube.com/watch?v=-Re_elg67e8&list=PLLoKvRqQVbtKJFde-bynhnXdePwI7jpS&index=2&t=4378s).
- [13] UNESCO. *UNESCO Recommendation on Open science. Programme and meeting document SC-PCB-SPP/2021/OS/UROS*, 2021. <https://unesdoc.unesco.org/ark:/48223/pf0000379949>
- [14] Vilela Rodrigues Rezende, Laura, e Sonia Barbosa. «Using the DATAVERSE project to move towards fair principles». In *Princípios FAIR aplicados à gestão de dados de pesquisa*, 2021. <https://www.arca.fiocruz.br/handle/icict/51281>.
- [15] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan 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. 1 (December 2016): 160018. <https://doi.org/10.1038/sdata.2016.18>

---

<sup>39</sup> Cfr. <https://www.coar-repositories.org/old-pages/activities/advocacy-leadership/working-group-next-generation-repositories/background/>

# ODI and BACODI: a study on *Destini incrociati* by Italo Calvino with Semantic Web Technologies

Enrica Bruno<sup>1</sup>, Valentina Pasqual<sup>2</sup>, Francesca Tomasi<sup>3</sup>

<sup>1</sup> University of Bologna, Italy – enrica.bruno@studio.unibo.it

<sup>2</sup> University of Bologna, Italy – valentina.pasqual2@unibo.it

<sup>3</sup> University of Bologna, Italy – francesca.tomasi@unibo.it

## ABSTRACT

On the centenary of Italo Calvino's birth, this research aims to investigate the “macchina narrativa combinatoria” [1] using Semantic Web technologies to formalise and analyse textual and iconographical aspects that concern narrative and compositional structures, plot concepts and character roles in the collection *Il castello dei destini incrociati* by Italo Calvino. In particular, ODI Ontology (Ontologia dei *Destini incrociati* di Italo Calvino) and its corresponding Knowledge Base (BACODI, Base di Conoscenza dell'Ontologia dei *Destini incrociati* di Italo Calvino), represent and describe the publication of the edition, the description of the cards - both as parts of the deck and as narrative instances in Calvino's text - and, finally, the semantic links and narrative relations between cards in each story, considering their coexistence in a single environment. Data has been analysed under three main aspects: the semantics conveyed by the tarot cards, the text structure and the relations between cards.

## KEYWORDS

Semantic Web, Italo Calvino, Narratology, Knowledge Graph, Digital Humanities

## 1. INTRODUCTION

*Il castello dei destini incrociati*, the first collection of the homonymous work published by Einaudi in 1973, is one of Calvino's most rigorous works of combinatorial literature. In particular, the author chooses a specific tarot deck, the *Pierpont-Morgan Bergamo* [2], as a system of signs and a real language [3]: the twelve stories of the collection arise, in fact, from the combination of several cards that, placed in a row - in vertical or horizontal order -, create a series of narrative sequences. The main characteristic of the travellers of *Il castello* is that they have lost their speech, but they are nevertheless eager to tell their own stories. It is this urge to communicate that leads them to use the tarot deck as a means of storytelling. The result is a *quadrato magico* [1] in which each narrative sequence is crossed with another, as are the destinies of the twelve travellers of the castle. Each sequence can be read in any direction (e.g. from top to bottom or from left to right and vice versa) so that the meaning of each narrative story is given by the way each card is combined with the other cards of the deck.

Although Digital Humanities research extensively covers knowledge representation of cultural heritage, there is a lack of formal representations for hermeneutic analysis of texts that consider formal expressions and narrative connections.

*Il castello dei destini incrociati*, because of its mechanical, geometric and combinatorial nature, lends itself to experience a formal ontology for a computational study both critical and narratological of the text. Specifically, ODI aims to represent the edition of the work, the description of the cards - both as artefacts and narrative instances - and the semantic links and narrative relations between the different cards in each story, considering their coexistence in a single environment. In summary, this research aims at superseding the canonical representation of textual entities (i.e., tagging places, agents and events of a text) modelling the texture of the work in all its narrative and compositional elements considering their textual and iconographical dimensions.

## 2. STATE OF THE ART

The field of digital scientific editions (DSE) has seen important developments, in the context of the semantic web, that have led to a new reflection on the concept of “digital edition” that goes beyond the reductive dichotomy based on medium (print-digital). Despite the widespread production of DSE as testified by [4, 5], to the best of our knowledge Calvino's production has received little attention in the field of Digital Humanities. *Atlante Calvino* [6]: *letteratura e visualizzazione* is a virtual environment that explores Calvino's corpus through a combination of literature and visual design but does not delve into its textual aspects. However, through an analysis of the textual elements, it could be feasible to decipher certain

portions of the texts that are still safeguarded by copyright, despite the inability to provide unrestricted access to them in their entirety.

Trying to build an ontology on *Il castello* by Italo Calvino means thinking, first, about its peculiar narrative and compositional structure.

Even though there are theoretical studies on *Il castello dei destini incrociati* (including both collections), there is an absence of analyses for a closer understanding of the combinatorial and narrative relations between cards in the text. In other words, many scholars pause over the multipurpose and symbolic nature of tarot cards for the generative process of the stories and their interweaving [15], but how the mechanism of the “macchina narrativa combinatoria” [1] works is overlooked. A computational study may facilitate in this regard a new attention to the combinatorial functioning of the work, through a study focused on the textures of the text and their characteristics.

Over time, various attempts have been made to identify a universal theoretical model of narrative structures, with increasing attention paid to the modalities and techniques of narration: the study and analysis of the character, starting with Propp's [7] famous scheme on character functions; reflections on the 'narrative voice' and the 'point of view' [8]; furthermore, the categories of space and time analysed by both [9] and [10].

Recently, attention has turned toward the use of ontological technologies for the representation of narrative structures for both literary and narratological research purposes. Some projects have attempted to integrate Semantic Web technologies and narratological studies as SEBNET [11], a system able to analyse and generate fairy tales; SEB [12,13], a semantic network application to analyse relations between characters in more complex narratives; Labyrinth [14], a system able to make users explore a digital archive following the narrative relations between the works of art contained in it.

### 3. METHODOLOGY

The source edition, particularly, the first six chapters of the first collection of Italo Calvino's *Il castello*, has been analysed according to the methodology proposed by [16]. This qualitative approach focuses on the text content and thematic analysis and consists of three main phases: *skimming* (which includes a preliminary, superficial analysis of the text), *reading* (thorough and precise analysis of the same), and *interpretation* (related to the moment of hermeneutic study of the work). The analysis produced 19 requirements (competency questions) grouped into 8 categories (e.g. “What are the meanings associated with cards that have suit Wands?” which concerns the category of the semiotics of suits in numeral cards, cf. section 5 (*Results*, a-h)).

Additionally, the analysis of source materials produced a series of conceptual maps representing the domain. Maps have been then refactored and translated into an OWL ontology, called ODI. Data have been gathered from the text and saved in a set of tables, which have been then converted into the Knowledge Base, called BACODI, modelled on the ontology. As a first testing activity, ODI has been applied to the second section of Calvino's work, *Tutte le altre storie* (seventh chapter, containing six stories) and the model validity was confirmed. Subsequently, even the second section of the work was added to the Knowledge Base. The representativeness of the ontology and the correctness of the data in BACODI have been tested through the 19 competency questions previously defined via SPARQL queries.

Finally, BACODI data have been studied to investigate the 8 main categories mainly concerning the work's semantics, the text structure and links and relations between cards. Each analysis is referred to a specific category and has been carried out using a selection of competency questions run in SPARQL against BACODI. In section 5 (*Results*) we summarise our findings, and we provide two example analyses (The semiotics of suits in numeral cards and iconographic dimensions of cards). The work is provided with technical documentation<sup>1</sup>, particularly, in *Query Paths* section all analyses are presented in detail.

### 4. MODEL

ODI models the requirements mentioned in section 1 (*Introduction*) formalising the edition of the work, the description of the cards - both as artefacts and narrative instances - and the semantic links and narrative relations between the different cards in each story. This section presents the model along with a case study, the representation of *The Cavaliere di Coppe* and its relation with another card (*La Temperanza*) in the first story.

---

<sup>1</sup> <https://odi-documentation.github.io/materials/>

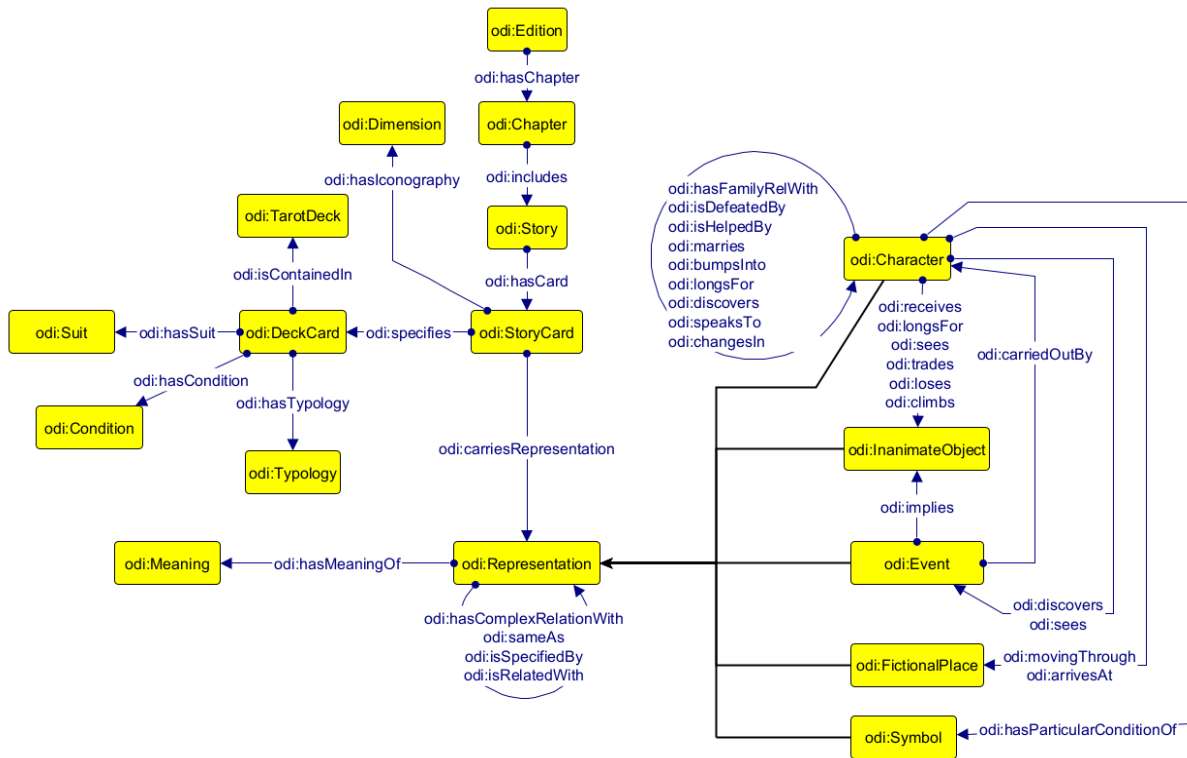


Figure 1. Main classes and object properties in ODI.

*The edition.* It represents the literary edition (`odi:Edition`), including title, author, publication date, publisher, place of publication and chapters. Although not directly addressed by the project’s representational requirements, ODI includes the description of the edition for both providing contextual information about the work and for further alignment with other ontologies.

*Tarot cards as artefacts and narrative instances.* It represents tarot cards (`odi:TarotCard`) as iconographical artworks belonging to the tarot deck *Pierpont-Morgan Bergamo* (`odi:TarotDeck`) and as occurrences in the stories (`odi:Story`). Each card is represented by its name, presence or not of the variant name in the collection, suit, typology, number, current location, condition, iconography and author. The connection between the card as a physical object (`odi:DeckCard`) and its narrative dimension (`odi:StoryCard`) is represented through the property `odi:specifies`. As specifications in the stories (`odi:StoryCard`), cards are, then, represented as narrative instances (`odi:Representation` and its subclasses, `odi:Meaning`) and with additional information such as the dimension of the iconography in the page, the position in the text and the presence or not of the variant name in the text.

*Semantics links and narrative relations between the cards in the stories.* It represents semantic links between two cards in the same story and the network of narrative relations between each protagonist and the cards in each story. As far as narrative relations are concerned, the model formalises simple relations (e.g. `odi:bumpsInto`, `odi:speaksTo`) and complex relations (`odi:hasComplexRelationWith`), which means direct or indirect relations. In particular, cards that have a direct relation with the protagonist can either be specified (`odi:isSpecifiedBy`) or have a general relation (`odi:isRelatedWith`) with the card that has got an indirect relation. In the Knowledge Base, complex relations are characterised by two triples, while simple relations by one. Two cards can, also, have a semantic link to express an identity (`odi:sameAs`) or an evolution (`odi:changesIn`) of the same representation in the same story.

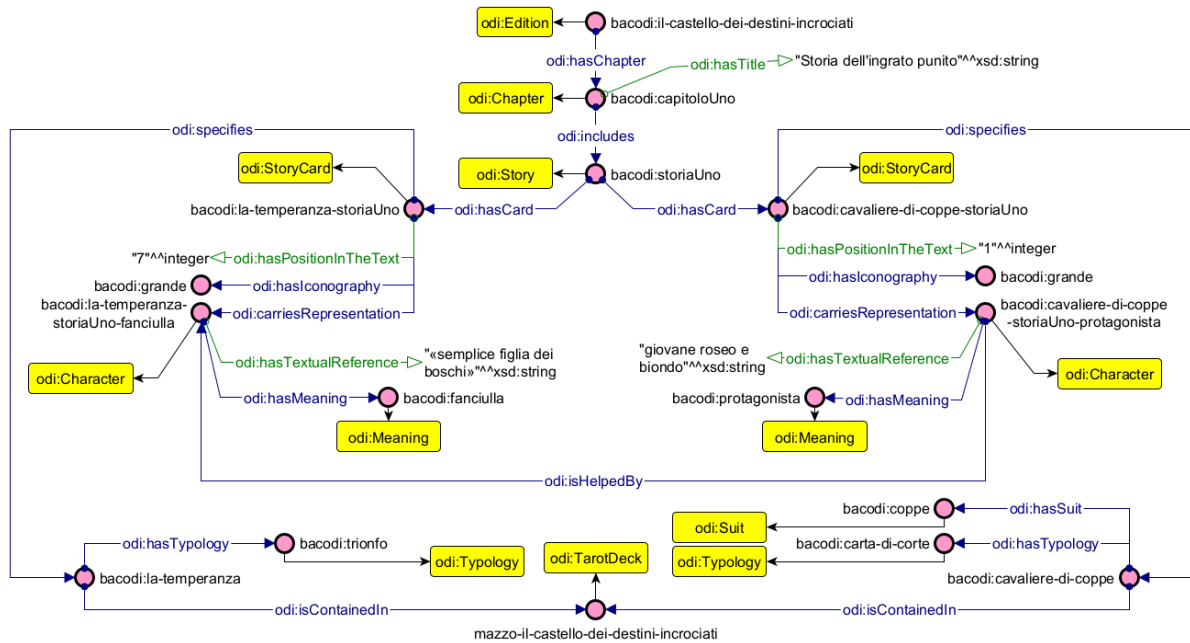


Figure 2. Representation in BACODI of the interaction (odi:isHelpedBy) between the cards Cavaliere di Coppe and La Temperanza in the first story.

Therefore, we present an example (figure 2), showing the formal representation of the *Cavaliere di Coppe* and its relation with another card (*La Temperanza*) in the first story. *Cavaliere di Coppe* (odi:StoryCard) is the protagonist of *Storia dell'ingrato punito* (odi:Story). After introducing himself, the protagonist arranges 17 tarot cards gradually. In summary, the narrative depicts a knight who is robbed and hanged from a tree branch by a brigand but subsequently rescued by a maiden. The said maiden, in turn, resolves to avenge the knight when he forsakes her to wed a woman of greater affluence. In the denouement of the story, the knight meets his demise at the hands of Cybele and her adherents. In particular, the maiden who initially helps the knight is represented by the triumph *La Temperanza* and in BACODI the property which describes this interaction is odi:isHelpedBy. As shown in figure 2, both cards are represented in relation to the tarot deck (odi:TarotDeck) and the story (odi:Story). In the first case, they are represented as artefacts belonging to the tarot deck (odi:TarotCard) with their main characteristics (e.g. the *Cavaliere di Coppe* is “cour card” (odi:Typology) and its suit is “cups” (odi:Suit)). In the second case, they are represented as narrative instances (odi:StoryCard) describing the interaction of the two characters in the story (odi:Character) with a specific meaning (respectively the protagonist and the woman, odi:Meaning).

## 5. RESULTS

The network of narrative occurrences preserved in BACODI and modelled using ODI, suggests the need for a more comprehensive analysis of the *Il Castello* considering the use of tarot cards as a privileged means of communication and, in particular, of narration. The analysis was on the whole carried out on three main aspects, recalling the categories of requirements (cf. section 1, *Introduction*): the **tarot cards semantics**, investigated through (a) the semiotics of suits in numeral cards, (b) the narrative function of court cards and triumphs [15], (c) the usage of the same cards with the same meaning in different stories, and (d) the use of two different cards for the same representation in the same story (semantic links, cf. section 4, *Model*); the **text structure**, investigated through (e) the position of the cards in each story, (f) recurring patterns of cards through the stories, and (g) the iconographic dimension of the cards running along the margins of the edition; finally, **relations between cards** have been analysed under the light of (h) simple relations (direct relations between each protagonist and the cards in the stories) and complex relations (indirect relations between each protagonist and the cards in the stories). For the sake of the example, here we report an analysis of the tarot cards' semantics (e.g., the semiotics of suits in numeral cards) and one of the text structure (the iconographic dimensions of the cards).

Each suit in *Il castello* has its own fictional iconography referring to a specific semantic field<sup>2</sup>, which differs from the traditional cartomantic bibliography [17]. The occurrences of tarot cards featuring the suit of Wands predominantly depict

<sup>2</sup> For the complete overview of this analysis please refer to the *Query Paths* page at <https://odi-documentation.github.io/materials/>. In particular, this study is documented in *The semantics > The semiotics of suits in numeral cards*.

the wooden environment of *Il Castello* (e.g., log, wood). However, in rare cases it is the card number that defines its meaning: for example, *Due di Bastoni* can be used to express duality or a choice between two situations (in *Storia della sposa dannata* and in the fourth story of *Tutte le altre storie*) regardless of the suit. In contrast, tarot cards featuring the suit of Coins refer to two distinctly contrasting semantic domains: first, earthly power, wealth and materiality (e.g., richness, inheritance, money); secondly, heaven, the divine and the immaterial (e.g., light, moon, state of beatitude). Between the two semantic fields, however, the former stands out over the latter. Most of the findings regarding the symbolism of numerals in the suit of Swords refer to themes of warfare (e.g., war, duel, dispute); sometimes, it specifically refers to defence (e.g., barrier, guards). Finally, the semantic meaning associated with the suit of Cups is less clearly defined than the three aforementioned suits. From the results, it is possible to note a large percentage that refers to banqueting and refreshment (e.g., banquet, wedding banquet, tavern); there are, however, other references such as those related to wealth (e.g., the source of life, world peace) or to love (e.g., love encounter). In most cases, the meanings of tarot cards are determined by their respective suits, with the type of suit significantly impacting the meaning of numbered cards (unlike triumphs and court cards, analysed in their narrative function (b)). Additionally, not all four suits refer to a single and defined semantic field: the suit of Wands recalls, in fact, a narrow semantic field, whereas the suit of Cups indicates broader semantics.

The existence of different iconographic dimensions in the text [1] has raised the prospect of investigating BACODI in order to comprehend the underlying motives for a decision that appears deliberate<sup>3</sup>. Overall, there are a total of 118 iconographic depictions running along the edition's margins, comprising 61 large and 57 small images.

The three queries made for the iconographic dimension of the cards show that, as far as numerals and triumphs are concerned, the discrepancy between the number of big dimensions and small dimensions is little.

Differently, the iconographic dimension of the court cards, and, particularly, those representing the protagonist, is almost always big except for three occurrences in the seventh chapter, *Tutte le altre storie*, in which it is drawn on the page with a small dimension (*Il Bagatto* in the fifth story; *La Temperanza* and *La Giustizia* in the last story). In all three occurrences, however, the three cards express something else than the presentation at the beginning of the story. In other words, the big dimension serves, at the beginning of the story, to the reader to identify the protagonist's card; the other times the protagonist appears with a different card, it can be either big or small. For instance, in the fifth story, the protagonist is, initially, represented with the *Fante di Coppe* and it has got a big dimension. After, the protagonist is represented with *Il Bagatto* to express the protagonist's desire to become emperor, and it has got a small dimension. The author's artistic choice to give importance to the dimension of each card in the text is thus proven. In the case of the protagonist card, the big dimension indicates the importance of the protagonist as the main motor of the narrative.

## 6. CONCLUSION AND FUTURE WORKS

This study explores how to exploit Semantic Web technologies to represent aspects of interpretative analysis in the humanities on literary texts. Specifically, this research aims to provide a formal representation of the narrative use of tarot cards and their relations in the stories *Il castello dei destini incrociati* by Italo Calvino through semantic web technologies. ODI and BACODI represent the textual and iconographical elements on which narratives are built (description of the cards - both in relation to the deck and to the stories in the collection) and their relations (the semantic links and narrative relations present in each story between the different cards) considering their coexistence in a single environment (the edition of the work).

The project supersedes the canonical representation of textual entities (i.e., tagging places, agents and events of a text) establishing a network of narrative relations and semantic links which allow to learn more about the construction of the combinatorial machine set up by the author. Thanks to the formalisation of data, it is possible to understand: the preference for some cards rather than others to express certain entities; how the cards are repeated and combined; what is the main structure of the various narrative sequences and how the literary tradition fits inside. Particularly, BACODI provides an example (cf. section 5, *Results*) of how modelling textual narrative structures as well as textual-iconographic relations can open further investigations on the "macchina narrativa combinatoria" [1] of *Il Castello*.

Results could encourage studies of other works of Calvino's combinatorial narrative period (1967-1980), adapting some ODI parts to the specificities of the texts. For instance, the unit including semantic links and narrative relations between cards in each story could be replicated to create a semantic network of all the characters appearing in the different readings of *Se una notte d'inverno un viaggiatore* (1979). This kind of attention could point out the role of the various characters in relation to the protagonist, the Reader, and to his plural reading path.

---

<sup>3</sup> For the complete overview of this analysis please refer to the Query Paths page at <https://odi-documentation.github.io/materials/>. In particular, this study is documented in *The text structure > The iconographic dimension*.



Moreover, this project represents the need to go beyond the canonical digital scholarly editions, even those based on semantic Web technologies to demonstrate how to use the semantics as a way for creating new interpretational paths, to explore the content of literary works and to expand, with new methodologies, the users' knowledge on our textual tradition. In the future, we aim to go beyond the approach that sees each protagonist as a privileged point of observation to formalise a more complex network of relations between all the cards in the deck. Additionally, we plan to align ODI with existing ontologies to ensure interoperability with other models (e.g., FRBRoo). We, also, aim to expand our work to the second collection of the work, *La taverna dei destini incrociati*. Finally, a future goal is to be able to create digital visualisations to explore the work through specific pathways between cards that can bridge the copyright issue. Particularly, for each card, a legend could be provided for the description and physical appearance of the card, with a related picture of the miniature; furthermore, different reading directions would accompany the reader within a plural path, according to the different narrative sequences in which each card may manifest itself.

## REFERENCES

- [1] Calvino, Italo. *Il castello dei destini incrociati*. Milano: Mondadori, 2014.
- [2] Kaplan, Stuart R. *The Encyclopedia Of Tarot*, vol 4. United States Games Systems, 1978. <https://archive.org/details/encyclopediaofta0000kapl/page/n3/mode/2up>.
- [3] Albani, Paolo. "La letteratura come gioco combinatorio." *Griselda. Il portale di Letteratura*. 2018. <https://site.unibo.it/griseldaonline/it/approfondimenti/paolo-albani-letteratura-gioco-combinatorio>.
- [4] Sahle, Patrick. 'A Catalog of Digital Scholarly Editions', Online catalogue, v.4.082. edited by Patrick Sahle et al., <https://www.digitale-edition.de/>.
- [5] Franzini, Greta, Melissa Terras, and Simon Mahony. 'A Catalogue of Digital Editions'. In *Digital Scholarly Editing*, edited by Matthew James Driscoll and Elena Pierazzo, 1st ed., 4:161–82. Theories and Practices. Open Book Publishers, 2016. <https://www.jstor.org/stable/j.ctt1fzhh6v.13>.
- [6] Elli, Tommaso. "Research through Visualization in Literary Criticism. Adaptations of data visualization to scholars' interpretive work", PhD diss., (Politecnico di Milano, 2022). <http://hdl.handle.net/10589/183038>.
- [7] Propp, Vladimir J., *Morfologia della fiaba*, a cura di Bravo G. L., Torino: Einaudi, 1966.
- [8] Bernardelli, Andrea, and Remo Cesarini. *Il testo narrativo*. Bologna: Il mulino, 2005.
- [9] Genette, Gérard. *Figure III. Discorso del racconto*. Torino: Einaudi, 1976.
- [10] Bakhtin, Michail. "Le forme del tempo e del cronotopo nel romanzo." In *Estetica e romanzo*, 231-405. 1979.
- [11] Gigliozzi, Giuseppe. "Codice, testo e interpretazione." *Studi di codifica e trattamento automatico di testi*. Ed. G. Gigliozzi. Roma: Bulzoni (1987): 65-84.
- [12] Gigliozzi, Giuseppe, and Sandra Giuliani. "Una parola che non dice nulla." *Le Novelle per un anno* (1993).
- [13] Gigliozzi, Giuseppe, and Sandra Giuliani. "La rete delle formiche. Un'applicazione di SebNet." *Fine della storia e storie senza fine*. Ed. C. Cazalé. Nanterre: Université Paris X–Nanterre (1993): 137-157.
- [14] Damiano, Rossana, and Lieto Antonio. "Ontological representations of narratives: a case study on stories and actions." *OPEN ACCESS SERIES IN INFORMATICS* (2013): 76-93. <http://drops.dagstuhl.de/opus/volltexte/2013/4149>
- [15] Corti, Maria. "Il gioco dei tarocchi come creazione di intrecci." *Il viaggio testuale*, Torino, Einaudi (1978): 169-84.
- [16] Bowen, Glenn A. "Document analysis as a qualitative research method." *Qualitative research journal* 9.2 (2009): 27-40. <https://doi.org/10.3316/QRJ0902027>
- [17] Berti, Giordano. *Storia dei tarocchi: verità e leggende sulle carte più misteriose del mondo*. Milano: Mondadori, 2007.

# L'edizione di un graphic novel medievale: il *Romanzo di Tristan* in München, BSB, Cgm 51 come laboratorio di codifica (Progetto: TRISDE 51)

Maria Adele Cipolla<sup>1</sup>, Anna Cappellotto<sup>2</sup>

<sup>1</sup> Università degli Studi di Verona, Italia – adele.cipolla@univr.it

<sup>2</sup> Università degli Studi di Verona, Italia – anna.cappellotto@univr.it

## PAROLE CHIAVE

Gottfried von Strassburg; *Tristan*; digital documentary edition; manoscritti illustrati

## 1. MÜNCHEN, BAYERISCHE STAATSBIBLIOTHEK, CGM 51

La prima metà del XIII secolo inaugura un nuovo *trend* nella cultura manoscritta di area tedesca, con la produzione di libri illustrati per le opere in volgare.<sup>1</sup> In questo contributo se ne prenderà in considerazione un caso-studio, il codice München, Bayerische Staatsbibliothek, Cgm 51:<sup>2</sup> il manoscritto, che risale al 1240 ca. per il testo poetico, ma è corredato di illustrazioni e iscrizioni realizzate fra la seconda metà del XIII e il XV secolo, rappresenta il testimone più antico del *Tristan* di Gottfried von Straßburg (1210 ca.) e della continuazione (*Fortsetzung*) di Ulrich von Türheim, ma ne tramanda una versione scorciata, priva di una rilevante quantità di versi in passaggi significativi della trama, e per questo è stato considerato dai filologi una redazione autonoma del testo (Marold 2004/1<sup>a</sup> ed.: 1906).

Allo stato attuale, il documento include 15 fogli illustrati a tutta pagina, forse concepiti indipendentemente e successivamente rilegati tra le carte che compongono il codice (Montag & Gichtel 1979). Le illustrazioni (dovute ad artisti diversi e di diversa qualità, appartenenti ad almeno due *workshop*, e spesso maldestramente ritoccate per ovviare al deterioramento delle immagini nel tempo) rappresentano una traduzione intersemiotica del testo poetico e sono corredate di iscrizioni vergate da più mani (nel corso di due secoli), collocate in rotoli e cartigli appositamente predisposti o secondariamente apposte in altri spazi delle illustrazioni. Queste iscrizioni corrispondono a varie tipologie testuali: se, infatti, nella maggior parte dei casi esse registrano i nomi dei personaggi raffigurati, talora trasmettono testi più ampi, riassumendo l'azione (didascalie narrative) o facendole da contrappunto, citando le battute dei personaggi come in un *graphic novel* (didascalie drammaturgiche), o dando voce al narratore onnisciente, che commenta l'azione con citazioni dotte o proverbi (didascalie parenetiche).

## 2. OBIETTIVI EDITORIALI

L'interazione tra elementi verbali e figurativi nell'esperienza dei fruitori contemporanei è uno degli aspetti caratterizzanti l'utilizzo dei manoscritti illustrati e, come ha osservato Nichols: “The dynamic of the medieval manuscript matrix – [...] particularly of illuminated manuscripts – involves cognitive perception as two kinds of literacy: reading text and interpreting visual signs” (Nichols 1990: 8). Al proposito, si è auspicata la fondazione di una “developmental theory of illustrated vernacular manuscripts” (Walworth 2007): nel Cgm 51, infatti, le illustrazioni ‘traducono’ i testi verbali e i diversi e successivi layer di iscrizioni circostanziano e ridefiniscono il senso delle illustrazioni e dei testi poetici, all'interno di una comunità ‘attiva’ di fruitori, che intervengono sul documento nel corso del tempo. Il codice attuale rappresenta quindi il prodotto multimodale e multiautoriale di una collettività di agenti, che contribuiscono progressivamente a costituire il manuscript matrix, che l'editore ha l'obiettivo di rendere accessibile attraverso l'edizione. Il filologo è chiamato quindi a implementare un'edizione che visualizzi la complessità del documento e la stratificazione dei suoi elementi costitutivi. L'obiettivo finale di questo progetto è dunque la creazione di una Digital Documentary Edition (DDE) (Pierazzo 2011), ossia un'edizione che si fondi nel documento e cerchi di renderlo accessibile all'utente, fornendo l'interpretazione

---

<sup>1</sup> Per una panoramica si veda il *Katalog der deutschsprachigen illustrierten Handschriften des Mittelalters* (<https://kdih.badw.de/das-projekt.html>, ultimo accesso: 25/4/2023), un progetto della Bayerische Akademie der Wissenschaft sul rapporto testo/immagine nel Medioevo di lingua tedesca e, all'interno del progetto, il database digitale di manoscritti illustrati (*Deutschsprachige illustrierte Handschriften des Mittelalters* <https://kdih.badw.de/datenbank/start>, ultimo accesso: 25/4/2023), organizzati secondo gruppi tematici (ad es. Dietrich von Bern) e sottogruppi, ovvero opere a cui si riferiscono (ad es. *Eckenlied*, *Laurin*, etc.), manoscritti e stampe.

<sup>2</sup> Il manoscritto è stato digitalizzato secondo il protocollo IIF ed è accessibile nel catalogo della Bayerische Staatsbibliothek <https://www.digitale-sammlungen.de/view/bsb00088332?page=1> (ultimo accesso: 25/04/2023).



critica delle sue singole parti e delle relazioni che tra esse intercorrono. Le illustrazioni e le iscrizioni contenute nel Cgm 51 non devono infatti essere pensate come elementi decorativi o accessori al testo poetico; sono al contrario elementi costitutivi del testo multimodale tra loro in continua interazione.

### 3. IL PROGETTO TRISDE 51

Il progetto TRISDE 51,<sup>3</sup> che ha preso avvio in seno al Progetto di Eccellenza 2018-2022 del Dipartimento di Lingue di Univr, si propone dunque di valorizzare quest'interpretazione degli elementi costitutivi del documento in un'edizione scientifica digitale, associando l'analisi critico-testuale all'interpretazione delle miniature e delle iscrizioni: in sostanza, l'obiettivo di ricerca preliminare è la valutazione di come la codifica in XML/TEI consenta di creare un modello di edizione in cui le diverse componenti dell'oggetto (testi – immagini – testi) possano essere analizzate e messe in relazione.

L'obiettivo finale consisterà nella creazione di un'edizione che può essere definita sperimentale: *in primis*, in ragione del lavoro di studio e di analisi di un manoscritto che integra testi e illustrazioni; in secondo luogo, per le sfide che pone la rappresentazione di un oggetto complesso attraverso i linguaggi di codifica e, infine, per le difficoltà che richiede la pubblicazione dell'edizione, stante la mancanza di un'applicazione adatta alla visualizzazione e la conseguente necessità di lavorare allo sviluppo di uno strumento *ad hoc*. Il progetto può rappresentare inoltre un contributo significativo alla teoria dell'edizione e, in particolare, all'ecdotica di codici illustrati, perché combina elementi di critica testuale (*textual criticism*) con elementi di critica visuale (*image criticism*).

Per il suo sviluppo, TRISDE 51 ha guardato ad altri progetti fondati su simili premesse e obiettivi, che potessero costituire un punto di partenza, ma si segnala che ad oggi non si è ancora lavorato sufficientemente alla creazione di questo tipo di edizioni, nonostante l'esistenza di molti documenti illustrati con una relazione stretta tra immagini e contenuto testuale (cfr. Cappellotto 2020). Sicuramente, come pietra miliare per edizioni del genere va citato il *Welscher Gast Digital* (WGD),<sup>4</sup> un progetto che si pone l'obiettivo studiare l'apparato iconografico 'seriale' nei testimoni dell'opera di Thomasin von Zerclaere, per individuare corrispondenze e divergenze nella rappresentazione visuale, e di realizzare una nuova edizione critica del testo. La differenza più consistente risiede nel fatto che TRISDE 51 prende in considerazione l'unico manoscritto del *Tristan* testimone di un simile apparato di illustrazioni, mentre il WGD considera l'intera tradizione di un ciclo iconografico pluritestimoniale, basato probabilmente su un archetipo autoriale.<sup>5</sup>

### 4. CODIFICA

Seguendo il *workflow* per la creazione di un'edizione, è stata innanzitutto condotta un'analisi della fonte. Il progetto si è posto come primo obiettivo l'individuazione degli elementi costitutivi dell'oggetto (immagini, testi) e della definizione delle relazioni tra le parti, attraverso la creazione di un vocabolario e di una grammatica degli elementi visuali e delle peculiarità della loro 'sintassi' nel documento, oltre alla categorizzazione dei testi delle iscrizioni, secondo una tassonomia definita di testi diegetici, drammaturgici, parenetici (Cipolla 2014 e 2023). La combinazione di scenografie, personaggi, oggetti, gesti e posture, tutti elementi convenzionali e recursivi, serve infatti a rinforzare o parodiare le norme comportamentali rappresentate nel testo poetico, mostrando inoltre la rilevanza della comunicazione non verbale nella società cortese di riferimento.

L'analisi della fonte con metodologie digitali si è focalizzata su aspetti materiali del manoscritto, come la struttura di fascicolazione, lo studio delle illustrazioni e l'analisi paleografica delle iscrizioni. Per rappresentare la struttura di collazione (Figura 1), rilevante poiché i fogli illustrati sono stati inseriti successivamente e capace di visualizzare le lacune determinatesi nel corso della tradizione, è stato utilizzato il software VCEditor,<sup>6</sup> a partire dal facsimile digitale messo a disposizione dalla Bayerische Staatsbibliothek secondo il protocollo IIF.<sup>7</sup>

<sup>33</sup> Si veda <https://dh.dlcs.univr.it/it/progetti/patrimonio-letterario-filologico/#trisinde51> (ultimo accesso: 25/04/2023). Il progetto ha beneficiato di un *grant* biennale dell'Università di Verona (Ricerca di Base 2017).

<sup>4</sup> <https://digi.ub.uni-heidelberg.de/wgd/> (ultimo accesso: 25/04/2023).

<sup>5</sup> Si veda <http://wgd.materiale-textkulturen.de/illustrationen/index.php> (ultimo accesso: 25/04/2023).

<sup>6</sup> <https://vceditor.library.upenn.edu/> (ultimo accesso: 25/04/2023). L'intera collazione con VCEditor è visibile a questo link: <https://vceditor.library.upenn.edu/project/632b0a11d459320001d1d64c/viewOnly> (ultimo accesso: 25/04/2023).

<sup>7</sup> Data la stratificazione cronologica delle iscrizioni, la fondatezza delle attribuzioni paleografiche potrà essere rafforzata dalle indagini multispettrali, al fine di una identificazione meno incerta delle mani intervenute nelle illustrazioni e nelle iscrizioni.

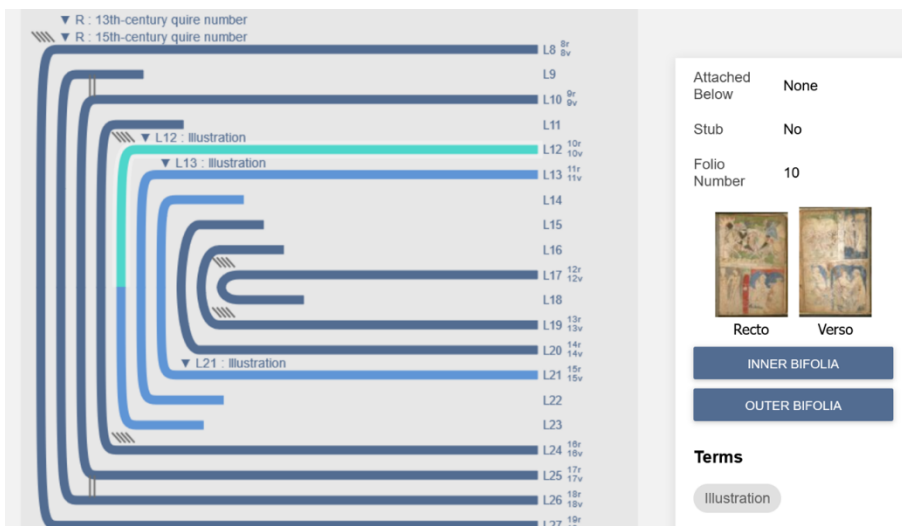


Figura 1. Esempio di visualizzazione del fascicolo 2 con VCEditor.

A questo fine e in seguito ai risultati dell'analisi della fonte, che hanno chiarito ulteriormente gli elementi rilevanti per gli obiettivi editoriali, sono state messe a frutto le possibilità offerte dalle linee guida TEI (cfr. TEI P5) e, in particolare, dai moduli per l'annotazione del facsimile digitale, per l'analisi interpretativa e per la trascrizione. Il lavoro di codifica del facsimile digitale è stato condotto sulla base dello studio preliminare (Cipolla 2014), in cui erano state proposte una tassonomia specifica per gli elementi costitutivi delle immagini, le tipologie testuali delle iscrizioni e una grammatica per la descrizione del rapporto tra testo e immagini. Sulla base di queste premesse, la domanda di ricerca ha riguardato la possibilità di rappresentare tassonomie e relazioni predeterminate attraverso il linguaggio di codifica e, in questo senso, TRISDE 51 è diventato un laboratorio di codifica.

A tale fine è stato necessario sviluppare un modello TEI complesso, innanzitutto avvalendosi del modulo per l'annotazione del facsimile.<sup>8</sup> Il modello concettuale seguito si è fondato su un procedimento che potremmo definire dal *macro* al *micro*: le superfici delle carte illustrate sono infatti suddivise in registri e in scene, al cui interno sono stati identificati gli elementi compositivi di rilievo (personaggi, oggetti di scena, gesti e posture) e i contenuti testuali (le iscrizioni). A ogni superficie o zona sono stati assegnati identificatori unici e altri attributi, utili ai fini dell'interpretazione o delle relazioni reciproche fra gli elementi.

L'annotazione del facsimile digitale ha suddiviso ogni carta illustrata (<surface>) prima nei registri di cui si compone (<zone>) e, all'interno dei singoli registri, nelle scene che costituiscono il *plot*, all'interno delle quali tutti gli elementi compositivi rilevanti sono stati annotati (Figure 2-4).

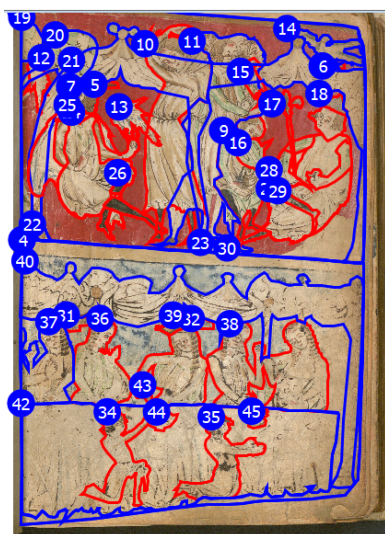


Figura 2. Esempio di annotazione del facsimile del f. 7r.

<sup>8</sup> Guidelines TEI P5, in particolare cap. 11: *Representation of Primary Sources*.

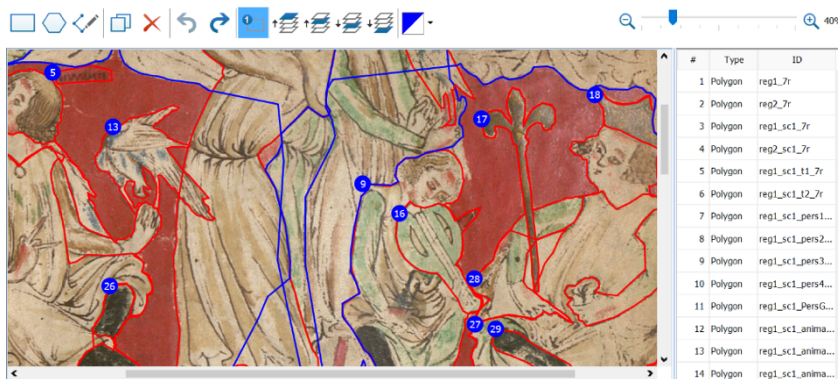


Figura 3. Esempio di annotazione del primo registro del f. 7r.



Figura 4. Esempio di annotazione di un registro con iscrizioni (f. 15v registro 2).

Agli elementi compositivi e ai testi presenti nelle illustrazioni sono stati assegnati ulteriori attributi, che rappresentano macrocategorie e categorie, identificate attraverso una serie di liste annotate predisposte all'interno del <teiHeader>.

Si è proceduto infatti alla creazione di una serie di gruppi interpretativi (<interpGrp type="X" xml:id="Y">) che contengono al loro interno tutte le categorie individuate, come nel seguente esempio (Figura 5), riferito al gruppo interpretativo 'animals', in cui si trovano tutte le tipologie di animali che compaiono nel ciclo iconografico:

```
<edition><interpGrp type="Animals" xml:id="animal">
  <desc>Animals</desc>
  <interp xml:id="bird">Bird</interp>
  <interp xml:id="bull_cow">Bull/Cow</interp>
  <interp xml:id="dog">Dog</interp>
  <interp xml:id="dragon">Dragon</interp>
  <interp xml:id="falcon">Falcon</interp>
  <interp xml:id="horse">Horse</interp>
  <interp xml:id="widdler">Widdler</interp>
  <interp xml:id="Petitcriu">Petitcriu</interp>
</interpGrp>
```

Figura 5. Gruppo interpretativo: animali.

I gruppi interpretativi includono quindi sia tutte le categorie che sono state individuate all'interno delle illustrazioni (che possono rappresentare posture, gesti, oggetti, persone, animali), sia le categorie testuali che classificano i contenuti delle iscrizioni a seconda della loro funzione (Figura 6), come nell'esempio seguente, dove vengono definite le tipologie testuali (ad es. prova di penna, testo parenetico, testo narrativo, etc.):

```

<interpGrp type="Text_types" xml:id="text">
  <desc>Text types</desc>
  <interp xml:id="pen_trial">Pen trial</interp>
  <interp xml:id="parenetic_text">Parenetic text (quotations, proverbs,
    narrator's comments)</interp>
  <interp xml:id="narrative_text">Narrative text</interp>
  <interp xml:id="name">Name</interp>
  <interp xml:id="speech">Speech (dramatic lines, monologues,
    dialogues)</interp>
  <interp xml:id="noun">Noun</interp>
  <interp xml:id="unreadable_text">Unreadable text</interp>
</interpGrp>

```

Figura 6. Gruppo interpretativo: tipologie testuali.

Anche gli episodi (Figura 7) che sono stati identificati all'interno delle illustrazioni e che corrispondono agli elementi del *plot* sono stati raggruppati all'interno di una macrocategoria 'episodes' che include tutte le scene individuate nell'analisi (ad es. festa di maggio, torneo, battaglia).

```

<interpGrp type="Episodes" xml:id="episode">
  <desc>Episodes</desc>
  <interp xml:id="May_Fest_Gentlemen">May Fest at Marke's court: gentlemen's
    pastimes: Cgm 51, f. 4rb 9-46; Ranke, 587-626</interp>
  <interp xml:id="May_Fest_Ladies">May Fest at Marke's court: ladies at
    banquet: Cgm 51, f. 4rb 9-46; Ranke, 587-626</interp>
  <interp xml:id="tournament_Ladies">Blanscheflur and Riwalin falling in love:
    ladies at the tournament: Cgm 51, f. 4va 26-5ra 16; Ranke,
    652-732</interp>
  <interp xml:id="tournament_Knights">Blanscheflur and Riwalin falling in
    love: knights at the tournament: Cgm 51, f. 4va 26-5ra 16; Ranke,
    652-732</interp>
  <interp xml:id="Riwalin_fight">Riwalin's fight against Marke's enemies: Cgm

```

Figura 7. Gruppo interpretativo: episodi.

In una fase successiva, il lavoro di codifica ha riguardato la trascrizione interpretativa dei testi presenti nelle illustrazioni: la trascrizione diplomatica e l'analisi paleografica delle mode scribali rappresentate dalle mani dei diversi agenti attivi nelle iscrizioni (i cui risultati sarebbero di estremo rilievo per la storia del manoscritto) è ancora in corso.<sup>9</sup>

La trascrizione segue il modulo per la rappresentazione delle fonti primarie e per la combinazione del facsimile con il testo mediante l'utilizzo dell'attributo @fac.<sup>10</sup> A tale proposito, sono utilizzati elementi come <pb> al cui interno sono stati inseriti divisioni e sottodivisioni strutturali che corrispondessero rispettivamente ai registri (<div1>) e alle scene (<div2>) e, all'interno di queste ultime, le trascrizioni sono state poste dentro l'elemento <ab>, dotato di attributi che riprendono la sottocategoria testuale (@ana), la porzione di facsimile corrispondente (@fac), oppure la lingua (@xml:lang) o lo stile (@style).

```

<pb n="15v" xml:id="f15v" facs="#fol15v"/>
<!-- TESTI REG 2 SC 1 -->
<div1 facs="#reg2_15v" type="reg" xml:id="f15vR2">
  <div2 facs="#reg2_sc1_15v" xml:id="f15vR2S1">
    <ab ana="#name" facs="#reg2_sc1_t1_15v" xml:lang="mhg" xml:id="f15vR2S1T1">
      Tristrant </ab>
    <ab style="poetry" ana="#parenetic_text" type="quotation"
      facs="#reg2_sc1_t2_15v" xml:lang="lat" xml:id="f15vR2S1T2"> Omnia que discis
      non aufert fur neque piscis </ab>
    <ab ana="#name" facs="#reg2_sc1_t3_15v" xml:lang="mhg" xml:id="f15vR2S1T3">
      Tristrant </ab>
    <ab ana="#noun" facs="#reg2_sc1_t4_15v" xml:lang="mhg" xml:id="f15vR2S1T4">
      Rotta </ab>

```

Figura 9. Esempio di trascrizione dei testi iscritti.

<sup>9</sup> Al riguardo, mancano ancora, purtroppo, strumenti di riferimento, quali progetti di analisi paleografica della documentazione tedesca del periodo, paragonabili, ad es., a DigiPal per la paleografia inglese del Medioevo.

<sup>10</sup> Cfr. TEI P5 11.2 *Combining Transcription with Facsimile*.

## 5. PROSPETTIVE

Chiarito il principale obiettivo editoriale, ovvero la creazione di una *Digital Documentary Edition* navigabile dal facsimile, partendo dal file di codifica TEI, abbiamo condotto alcuni esperimenti, tuttora in corso, sulla visualizzazione dell'edizione, per definire le funzionalità di un'applicazione *ad hoc* in cui l'utente possa condurre ricerche all'interno di una singola carta illustrata o navigando nell'intera collezione di fogli illustrati. Le liste analitiche predefinite nel file TEI suddivise per macrocategorie diventeranno indici selezionabili di tipologie di *realia*, personaggi, episodi, gesti, posture, testi: una volta individuata la macrocategoria e l'oggetto di interesse (ad es. personaggi/Tristano), l'applicazione visualizzerà le carte e i *loci* che in cui essi sono collocati. I testi verranno a loro volta visualizzati in trascrizione nel luogo in cui essi appaiono all'interno del manoscritto e verranno corredati di note critiche. Come già rilevato, l'applicazione è ad oggi in via di sviluppo, ma confidiamo che, già in questa fase, il lavoro possa rappresentare un contributo alla pratica delle edizioni scientifiche digitali di documenti manoscritti illustrati.

## 6. RINGRAZIAMENTI

Al progetto hanno collaborato Giulia D'Agostino (che ha contribuito allo sviluppo del modello di codifica), Lorenzo Ferroni (che si è occupato della visualizzazione della collazione con VCEditor) e Elisabetta Triolo (che ha sperimentato soluzioni per la visualizzazione dell'edizione). Un ringraziamento speciale va a Tiziana Mancinelli, titolare del corso di Informatica per gli Studi Umanistici all'Università di Verona nell'anno 2016/2017, che ha collaborato con il corso di Filologia Germanica, introducendo gli studenti allo studio del surrogato digitale del manoscritto e alla rappresentazione testuale attraverso la codifica TEI.

## BIBLIOGRAFIA

- [1] Baisch, Martin. «Wertlose Zeugen?: Formen von Materialität im Spannungsfeld von Textkritik und Kulturwissenschaft». In *Materialität in der Editionswissenschaft*. De Gruyter, 2010: 251-266. <https://doi.org/10.1515/9783110231311.251>
- [2] Becker, Peter Jörg. *Handschriften und Frühdrucke mittelhochdeutscher Epen: Eneide, Tristrant, Tristan, Erec, Iwein, Parzival, Willehalm, Jüngerer Titurel, Nibelungenlied u. ihre Reproduktion u. Rezeption im späteren Mittelalter u. in d. frühen Neuzeit*. Reichert, 1977.
- [3] Cappellotto, Anna. «From codex to apps: the medieval manuscript in the age of its digital reproduction». *Umanistica Digitale* 4, 9 (2020): 1-18, <https://doi.org/10.6092/issn.2532-8816/11459>
- [4] Cipolla, Adele. «Tristan in Munich, BSB, Cgm 5. Observations on an Intersemiotic Romance». In Bertagnolli, Davide & Alessandro Zironi (eds), *Fragmentation and Inclusion: Medieval Translation Inbetween*. Brepols, 2023 (in corso di stampa)
- [5] Cipolla, Adele. *Gli amanti nella selva. Herr Tristrant - BSB Cgm 51*. Fiorini, 2014.
- [6] D'Elden, Stephanie Cain van. *Tristan and Isolde: medieval illustrations of the verse romances*. Brepols, 2016.
- [7] Drecoll, Uta. *Tod in der Liebe, Liebe im Tod: Untersuchungen zu Wolframs Titurel und Gottfrieds Tristan in Wort und Bild*. Peter Lang, 2000.
- [8] Eming, Jutta, Rasmussen, Ann Marie & Kathryn Starkey (Hgg.). *Visuality and materiality in the story of Tristan and Isolde*. University of Notre Dame Press, 2012.
- [9] KdiH: *Katalog der deutschsprachigen illustrierten Handschriften des Mittelalters*, <https://kdiH.badw.de/en/kdiH-digital.html>.
- [10] Marold, Karl (Hg.). *Gottfried von Straßburg, Tristan*. Bd. 1: Text (Unveränd. 5. Abdr. nach dem 3., mit einem auf Grund von Friedrich Rankes Kollationen verb. kritischen Apparat besorgt und mit einem erw. Nachw. versehen); Bd. 2: Übersetzung. De Gruyter, 2004 [1a ed. 1906].
- [11] Nichols, Stephen G. «Introduction: Philology in a Manuscript Culture». *Speculum* 65, 1 (1990): 1-10, <https://doi.org/10.2307/2864468>
- [12] Ott, Norbert H. «Katalog der Tristan-Bildzeugnisse». In Frühmorgen-Voss, Hella (Hg.). *Text und Illustration im Mittelalter: Aufsätze zu d. Wechselbeziehungen zwischen Literatur u. bildender Kunst*. Münchener Texte und Untersuchungen zur deutschen Literatur des Mittelalters, Bd. 50. Beck, 1975: 140-171.
- [13] P5: Guidelines for Electronic Text Encoding and Interchange, TEI Guidelines Version 4.6.0. Last updated on 4th April 2023, <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>
- [14] Pierazzo, Elena. «A rationale of digital documentary editions», *Literary and Linguistic Computing* 26, 4 (2011): 463-477, <https://academic.oup.com/dsh/article/26/4/463/1055470>
- [15] Starkey, Kathryn & Wenzel, Horst (eds). *Visual culture and the German Middle Ages*. Palgrave Macmillan, 2005.
- [16] Ulrich Montag & Gichtel, Paul (Hgg.). *Gottfried von Straßburg, Tristan und Isolde: Faksimile Ausgabe des Cgm 51 der Bayerischen Staatsbibliothek*. Müller und Schindler, 1979.
- [17] VCEditor: <https://vceditor.library.upenn.edu/>

[18] Walworth, Julia. *Parallel Narratives: Function and Form in the Munich Illustrated Manuscripts of 'Tristan' and 'Willehelm von Orlens'*. Boydell & Brewer, 2007 (King's College London Medieval Studies 20).

[19] WGD - Thomasin von Zerklære: *Welscher Gast*. Text-Bild-Edition ›Welscher Gast digital‹. Hrsg. von Jakub Šimek unter Mitwirkung von Peter Schmidt und Christian Schneider. Heidelberg Universitätsbibliothek, 2015, <https://doi.org/10.11588/edition.wgd>.



# I Capricci d'un Re digitale

Elisa Conti<sup>1</sup>

<sup>1</sup> Università degli Studi di Catania, Italia – econti272@gmail.com

## ABSTRACT

Il presente contributo descrive le fasi di realizzazione dell'edizione digitale de *I Capricci d'un Re* di Stefano Pulvirenti. L'intento del progetto è stato quello di riportare alla luce e alla memoria un testo del Novecento andato perduto nel corso degli anni, proponendo una nuova edizione che ha come suo centro il manoscritto dell'opera. Il ritrovamento delle carte di Pulvirenti e la codifica del testo hanno dato la possibilità di produrre una nuova edizione che superi la mera trascrizione testuale. Il progetto viene proposto alla comunità tutta per mezzo del sito web da noi creato e dedicato all'opera suddetta e al suo autore.

## PAROLE CHIAVE

Edizione digitale, evt, codifica, manoscritto, Stefano Pulvirenti

## 1. INTRODUZIONE

L'idea da cui nasce un'edizione digitale è sempre legata alle innumerevoli e allettanti possibilità che offre oggi l'informatica umanistica nell'andare oltre la pagina stampata[7]. In questo caso si è voluto prediligere una tipologia di edizione che riesca a mettere il lettore quasi a contatto diretto con il testo manoscritto di Pulvirenti, grazie al software EVT<sup>1</sup>. Questo strumento di visualizzazione permette infatti di mostrare il facsimile del manoscritto e la pagina codificata contemporaneamente, in modo da offrire una lettura *trasparente*, totalmente fedele alle carte dell'autore. In questo modo il lettore stesso potrà egli osservare la grafia dell'autore, la fattura del quaderno, gli errori del testo [6]. Nonostante si perda l'esperienza tattile dello sfogliare il manoscritto o non si sia direttamente esposti dell'*aura* di cui parla Walter Benjamin[1], grazie ad una visualizzazione di questo tipo è possibile vivere un'esperienza maggiormente autentica rispetto alla lettura della sola trascrizione del testo che non avrebbe nulla di differente rispetto ad una classica edizione a stampa. Il lettore potrà sfogliare virtualmente il testo, utilizzando la lente di ingrandimento per i dettagli più minuti e avendo una visione d'insieme di tutto il manoscritto grazie alle miniature, i quali sono strumenti di EVT.

Il proposito è dunque di portare il lettore dentro il testo e sulla sua superficie materiale intrisa di un inchiostro che può egli stesso osservare.

## 2. I CAPRICCI D'UN RE DIGITALE

*I Capricci d'un Re* è una commedia in cinque atti di Stefano Pulvirenti, medico e scrittore siciliano vissuto a cavallo tra Ottocento e Novecento. Il testo viene scritto nei primi anni del XX secolo e pubblicato nel 1930 presso la tipografia romana Tipografia Consiglio Nazionale di Emigrazione e Lavoro. Non abbiamo dati sulla diffusione (o mancata disseminazione) del testo a stampa, di cui è registrata solo una copia presso la Biblioteca Nazionale Centrale di Firenze. Tuttavia, il ritrovamento del manoscritto nella biblioteca degli eredi di Pulvirenti permette oggi di riportare alla memoria un testo che potrebbe aggiungere valore al panorama teatrale siciliano del Novecento.

Il primo passo è stato quello di esaminare il testo del manoscritto confrontandolo con l'edizione del 1930. Dal confronto è emerso che il testo delle carte di Pulvirenti presenta pochissime varianti ortografiche rispetto al testo a stampa, dunque si potrebbe trattare della copia *in bella* affidata alla tipografia per la pubblicazione. A ciò si aggiunge il fatto che il manoscritto presenta pochissimi errori di cancellatura.

Una volta esaminato il contenuto, è iniziata la fase relativa alla modellizzazione dell'edizione con la scelta del tipo di codifica e dello strumento di visualizzazione da impiegare. La codifica scelta è XML-TEI e lo strumento di visualizzazione è EVT1.3 La fase di annotazione e di formalizzazione è stata fondamentale non solo per il prodotto che restituito ma soprattutto per l'atto ermeneutico ed ecdotico che essa richiede[4]. La codifica impone una riflessione critica sul testo trattato e una sua interpretazione che viene poi formalizzata nel linguaggio di marcatura utilizzato. Per *I Capricci d'un Re* digitale, la riflessione critica ci ha condotti verso la scelta di privilegiare il set di marcatori specifici del testo drammatico di cui parleremo successivamente.

---

<sup>1</sup> <http://evt.labcd.unipi.it/>

Superate le fasi preliminari, dunque una volta decise le caratteristiche dell'edizione, è iniziata l'operazione di scansione delle immagini. La rappresentazione rimane un elemento centrale delle edizioni, senza cui non si potrebbe avere il valore aggiunto dato dal digitale[7]. Questa implica l'operazione di cattura dell'oggetto culturale sia per i produttori dell'edizione che per i lettori. Per i primi si concretizza attivamente nell'atto della rappresentazione dell'oggetto, in questo caso del manoscritto, tramite le scansioni e la lavorazione delle immagini; per i secondi avviene grazie alla possibilità di fruire di tali immagini e poterle osservare e analizzare come se l'oggetto concreto fosse tra le loro mani. Secondo questa idea, sono stati prodotti vari formati in modo da fruire delle varie possibilità date da EVT (standard, lente di ingrandimento e miniature), e poter così osservare l'immagine nella sua interezza e nel dettaglio.

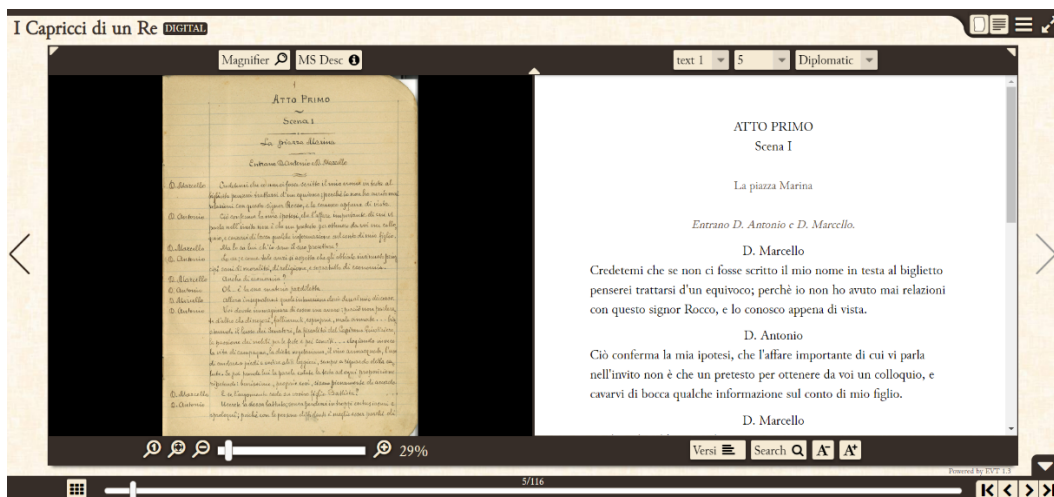


Figura 1. Esempio di visualizzazione del manoscritto con EVT1.

Successivamente è iniziata l'attività di codifica. Sono stati utilizzati i tag specifici del testo drammatico della TEI[8], in particolare per la lista dei personaggi sono stati utilizzati i marcatori <castList>, <castItem>, <castGroup> e il tag <roleDesc> per il ruolo dei personaggi. Ogni personaggio ha un xml:id che è servito nella fase di attribuzione delle battute. Alcuni personaggi sono stati inseriti da noi nella lista principale per mezzo di note in cui viene segnalato il nostro intervento. Per quanto riguarda le battute, è stato utilizzato il marcatore <sp> con gli altri due tag richiesti <speaker> e <p>. All'interno di <sp> è stato utilizzato l'attributo @who che ci ha permesso di inserire il rimando all'xml:id del personaggio. Proprio per questo motivo la lista dei personaggi iniziale presenta degli elementi aggiuntivi segnalati nelle note. Infatti l'autore ha inserito nuovi personaggi solo in un secondo momento, senza apportare modifiche a tale lista e dunque, per dare omogeneità alla codifica, abbiamo preferito inserire nelle note i personaggi aggiunti successivamente da Pulvirenti, esplicitando il nostro intervento negli appositi luoghi.

```

I Capricci di un Re.xml X
TEI teiHeader fileDesc sourceDesc msDesc physDesc objectDesc p
166 <pb n="2" xml:id="ICDR2"/>
167 <head>C.9<lb/>20/13</head>
168 </div>
169 <div xml:id="Lista_personaggi">
170 <pb n="3" xml:id="ICDR3"/>
171 <head>I capricci d'un Re<lb/>Commedia in cinque atti<lb/>Personaggi</head>
172 <castList>
173 <castItem type="role">
174 <role xml:id="Re">Carlo III,</role>
175 <roleDesc>Re delle Due Sicilie.</roleDesc>
176 </castItem>
177 <castGroup>
178 <castItem type="role">
179 <role xml:id="Ruffo">Principe Ruffo</role>
180 </castItem>
181 <castItem type="role">
182 <role xml:id="Eboli">Duca di Eboli</role>
183 </castItem>
184 <castItem type="role">
185 <role xml:id="Squillace">Marchese Squillace</role>
186 </castItem>

```

Figura 2. Lista dei personaggi codificata con OxygenXML Editor<sup>2</sup>.

<sup>2</sup> [https://www.oxygenxml.com/xml\\_editor.html](https://www.oxygenxml.com/xml_editor.html)



Per le entrate, le uscite dei personaggi, le attività sul palco e il modo di parlare è stato impiegato il tag <stage>. Grazie alla varietà di attributi (@entrance, @exit, @business, @delivery, @mixed, @modifier) è stato possibile distinguere le attività dei personaggi utilizzando un solo marcatore.

È stato scelto un tipo di codifica anche per le parole in una lingua diversa dall'italiano o per i termini sottolineati dall'autore (quasi sempre si tratta di termini siciliani). Oltre alla lingua italiana sono presenti parole in francese, spagnolo, latino e, come già detto, in dialetto siciliano. Quindi abbiamo evidenziato questi casi con il tag <foreign> con attributi @rend e @xml:lang nei casi delle parole in lingua straniera.

Infine, l'ultima operazione è stata quella di creare il sito dedicato ad ospitare l'edizione EVT e la sua analisi. Creato tramite CMS *WordPress*, contiene la biografia dell'autore e la descrizione del progetto. L'eventuale sviluppo futuro sarà quello di arricchire il sito con le edizioni dei vari testi prodotti da Stefano Pulvrenti (*Le zanzare e Trattato di Fisica Medicina. Alcune considerazioni sulla tecnica della Percussione*), in modo da riportare alla memoria collettiva il nostro autore siciliano e le sue opere.

## BIBLIOGRAFIA

- [1] Benjamin, Walter. *L'opera d'arte nell'epoca della sua riproducibilità tecnica*. Torino: Giulio Einaudi editore, 1966: 22-23.
- [2] Ciotti, Fabio. «Cosa è la codifica informatica dei testi». *Atti del Convegno Umanesimo & Informatica* (Trento 1996). Trento: Università degli Studi di Trento, 1996.
- [3] Del Turco, Roberto Rosselli. «Designing an Advanced Software Tool for Digital Scholarly Editions: The Inception and Development of EVT (Edition Visualization Technology)». *Textual Cultures* 12, n. 2 (2019): 91–111. <https://www.jstor.org/stable/26821538>
- [4] Italia, Paola, e Tomasi, Francesca. «Filologia digitale. Fra teoria, metodologia e tecnica». *Ecdotica* 11, n. 1 (2014): 112–130. doi: 10.7385/99218, <https://www.rivisteweb.it/doi/10.7385/99218>
- [5] Pierazzo, Elena. «Modelling Digital Scholarly Editing: From Plato to Heraclitus» in *Digital scholarly editing: Theories and practices*, 41-58. Cambridge: OpenBook Publishers, 2016. <http://books.openedition.org/obp/3399>
- [6] Pirandello nazionale, *Il fu Mattia Pascal*: <https://www.pirandellonazionale.it/download/evtms/>
- [7] Sahle, Patrick. «What is a scholarly digital edition?». *Digital scholarly editing: Theories and practices*, 19–39. Cambridge: OpenBook Publishers, 2016. <http://books.openedition.org/obp/3397>
- [8] TEI Consortium. «Core Tags For Drama – TEI: Text Encoding Initiative». TEI - Text Encoding Initiative. <https://www.tei-c.org/release/doc/tei-p5-doc/it/html/CO.html#CODR>.

# Harmonizing and modelling a bibliographic ontology of imperial diplomas within the OntoVE Knowledge Base

Chiara De Bastiani<sup>1</sup>

<sup>1</sup> University Ca' Foscari of Venice, Italy – chiara.debastiani@unive.it

## ABSTRACT

In this paper, I discuss the modelling of an ontological module for imperial diplomas relating to Veneto territories with the goal of adding value to the resources, by formalizing the contents to make the relationship between the diplomas and Veneto searchable through a dedicated User Interface. The harmonization of different bibliographic ontological models lies at the centre of our practice, and will enable a better description of the entities, thereby enhancing the subsequent interrogations of the ontology. The modelling is guided by the Knowledge Base competency questions, the nature of the sources and the context of production. The modelling approach described here aims to harmonize already established ontological models within the LOD community, such as FRBRoo and Bibframe. I will argue that this approach has several advantages, since it allows to keep CIDOC-CRM event-centric model through FRBRoo classes, and to enrich the module with finer-grained classes and properties as described in Bibframe, which also constitutes a standardized vocabulary for bibliographic description. I will propose a modelling of the contents of the diplomas as a performative act, to which information about the actors involved and the places is attached. In order for the users to locate the relevant diplomas, also the places mentioned in the diplomas are modelled as subjects of it. I will argue that reusing already existing ontologies and harmonizing them, rather than creating *ad hoc* classes, constitutes a more economic practice, which enhances interoperability of the module with already established standards.

## KEYWORDS

Imperial diplomas; bibliographic ontology; reusability; interoperability.

## 1. INTRODUCTION

With this paper, I discuss the modelling of an ontological module of written records represented by imperial diplomas starting from the Carolingian Empire and regarding Veneto territories. The ontological module is embedded within a wider project with the goal of creating a knowledge base of Germanic Cultural Heritage items in the Veneto region, the OntoVE Knowledge Base, to highlight the relationship between different Germanic populations and the Veneto territories. The KB will result from the close collaboration between academic scholars and a partner non-academic IT institution, Net7; the methodological choices, with their testing and implementation phases, result therefore from both scholarly assessment and technical expertise.

The modelling is guided by the nature of the sources, their context of production and the competency questions underlying the KB under construction. Starting from the latter, the competency questions underlying our KB should enable scholars to investigate, among other aspects, the complex relationship between Germanic populations and the Veneto territories. As an example, a user interrogating the Knowledge Base should be able to retrieve answers to the following questions: In which area are cultural heritage objects pertaining to the cultural scope x mostly found? Which written sources can be found for the chronologic period A-B? In which period or in which area is written production mostly found? Which kind of cultural heritage artifacts prevail for a certain time frame?

The relationship between Germanic populations and the Veneto territories is testified both by archaeological findings, which provide information about which areas of Veneto the first settlements were located, and help gather insights about their strategic importance [11], and by imperial diplomas, for instance. The dataset included in the OntoVE Knowledge Base is therefore both specific, since a specific portion of real-world knowledge is being described, and diverse, since the items included are of heterogeneous nature. Two ontological modules are devised to capture the items of the KB. The description of archaeological findings is carried out through the OntoVE\_Archeo module, which will communicate with the module devised for written records. In this paper, I describe the methodological choices underlying the semantic description of imperial diplomas.

Imperial diplomas, starting from Charlemagne after the conquest of the Lombard Kingdom, offer insights into the relationship between rulers of Germanic origin and the Veneto territories subjected to them. With respect to a time when Charlemagne was reorganizing the Empire, diplomas have a special position in witnessing different administrative acts

that help scholars gain insights into the relationship between the new Carolingian empire and the overthrown Lombard rulers and their successors.

## 2. MODELLING PRACTICE

Our modelling practice has as its centre the sources we are going to link in an Open Data environment; we decided to include collections that are already digitized and are openly accessible, in order to give scholars using our knowledge base and SPARQL endpoint the possibility of freely consulting the materials. The main source of data collection is the Digital *Monumenta Germaniae Historica* (dMGH, <https://www.dmg.de/>), which consists in the digitization of the volumes edited in the course of the 19th and 20th centuries and curated by the *Monumenta Germaniae Historica* Institution. The digitized collection constitutes a step forward in the open circulation of documents and knowledge, but suffers from some limitations. By browsing the dMGH, for instance, one has the possibility of searching for text-strings within the whole collection, so one could look for the string *Venetia\** and receive as results the portions of text in which it is found ([https://www.dmg.de/search?q=Venetia\\*](https://www.dmg.de/search?q=Venetia*)). The results obtained from this kind of search, however, need further manual filtering out in order to locate the documents relevant to one's research questions. Such limits are well known in the research on Digital Humanities, cf. [2], and recent projects, such as MeMO [2] or STOLE [1] among others, show how Semantic Web technologies can overcome them by providing formalized descriptions of the dataset chosen, and by making the data searchable through dedicated SPARQL endpoints. Dedicated user interfaces can then provide the user with analytical tools for further inspection of the documents. I argue therefore that devising an ontology of such documents will add further value to the resources by providing formalized descriptions of their contents and context of production.

A first distinction in the model is made between the conceptual form of a written document and its different redactions. Such distinction is not new in the field of bibliographic ontologies, since different standards offer a version of this bipartition, in some cases articulated in subtler divisions, as shown in [12].

In particular, the diplomas are formalized in the redaction provided in the editions of the dMGH; the nature of the source constitutes a further factor in the modelling practice: in these editions, the contents of the diplomas are transcribed from various sources, listed in the introductory section preceding the diploma itself. The introductory section also contains claims about alleged falsification of the whole or parts of a given diploma. This kind of information is also relevant and modelled in our ontological module.

As is evident from the nature of the sources to be included in the ontology, therefore, the basic conceptual distinction between the *work* (i.e. the contents of a written document) and its *redaction* is relevant for our modelling. This distinction is further expanded through the models reused in the present ontology.

In the light of the best practices in Digital Humanities, the OntoVE module reuses established formal ontologies; in particular, we are reusing the model Functional Requirements for Bibliographic Records object oriented, FRBRoo, v. 2.4,<sup>1</sup> [4], since this merges CIDOC-CRM's [3] Event-Centric model with the formal division proposed in the FRBR ontology between *Work*, *Expression*, *Item* and *Manifestation* [4]. There are several advantages in reusing FRBRoo; as [7] writes, FRBRoo not only conceptualizes the characteristics of a work in its different forms, as the FRBR does, but also formalizes the events around the production of a certain document; moreover, it is more suitable for the resource description framework data storage since it is formalized according to the CIDOC-CRM model.

Furthermore, the adoption of this formal ontology within the OntoVE KB enables interoperability at the level of the knowledge base, since CIDOC-CRM is reused in the OntoVE\_Archeo module mentioned in the Introduction; the FRBRoo model is in fact already aligned with CIDOC-CRM and reuses some of its properties and classes. As noted in the modelling of the OntoVE\_Archeo module, the CIDOC-CRM and the FRBRoo models provide abstract and general classes and properties, which have the advantage of enabling the description of very diverse items, but are resistant to concrete implementation [2]. For this reason, classes and properties from finer-grained ontologies are added to the ontological modules of the KB. In the module under discussion, we are reusing the Bibframe ontology, which constitutes moreover a standardized vocabulary for bibliographic description and allows for finer-grained descriptions of the items and the events

---

<sup>1</sup> A reviewer remarks that FRBRoo has been superseded by LRMoo in 2020; it must be noticed, however, that the LRMoo releases listed at [https://cidoc-crm.org/frbroo/fm\\_releases](https://cidoc-crm.org/frbroo/fm_releases) (last accessed 01/05/2023) are all still in draft form, and do not therefore represent a stable version of the model, such as the November 2015 FRBRoo release, approved by IFLA, used in our project. As stated in the release page, a draft version "is subject to issues identified and discussed on the SIG mailing list. This release should be used only for the purpose of following ongoing modeling discussions, or ones of historical interest. This document is not meant to support implementations, referencing or other official activities." ([https://cidoc-crm.org/frbroo/fm\\_releases](https://cidoc-crm.org/frbroo/fm_releases), status). For this reason, the stable version of the model has been included in the present ontological design, since a stable version can be used for implementation and referencing. Once LRMoo provides a stable version of the model, the relevant changes will be carried out in our ontological design.

around the production of diplomas.<sup>2</sup> The two bibliographic models are then harmonized, i.e. their classes and properties are aligned within our ontological module, created with the editor Protégé Desktop, version 5.5.0 (<http://protege.stanford.edu>), [9]. More specifically, classes that are semantically overlapping are aligned by means of an assertion of equivalence, whereas finer-grained distinctions provided by the Bibframe model are assigned to classes or properties of the FRBRoo model.

An example of the modelling of a diploma is provided in Figure 1 (cf. below): the prefixes *frbroo*, *bf* and *cidoc-crm* refer to the FRBRoo, Bibframe and CIDOC-CRM classes and properties respectively. Dotted arrows indicate the individuals, whereas full arrows represent properties. Notice that in Figure 1 we are focussing on the modelling and harmonization of the most relevant classes, therefore further properties related to classes not relevant for the present discussion are not represented.

In a nutshell, the model proposed in FRBRoo separates the abstract work from its expression (e.g. a translation in a different language), the items conveying the contents of a work (e.g. printed or electronic editions of the abstract work) and the manifestation (e.g. a unique copy of a document conveying the abstract work). Thanks to this first modelling, the redaction of a work as it is conveyed in the print edition and derived digitized dMGH form can be separated from the contents of the work itself. The different layers convey therefore different kinds of information: for instance, information about the actors involved, the original places and dates of conception of the work, and the subjects of the work are attached to the top layer conceptualizing the abstract work (cf. Figure 1 below).

A further methodological step involved in the modelling of this ontological module regards the alignment of the Bibframe and the FRBRoo models; Bibframe [8] provides a rich and standardized vocabulary for modelling different items according to MARC 21 standards. In its original conception, this model foresaw only two formal layers distinguishing between the *Creative Work* and its *Instance*, i.e. between the conceptual nature of a work and its physical embodiment on a carrier, cf [12]. However, the current version of the Bibframe model consists of three layers: *Work*, *Instance* and *Item*. The first two classes are mapped in our module to the FRBRoo classes *F1 Work* and *F3 Manifestation Product Type* respectively. The finer-grained conceptualization offered in Bibframe allows to explicitly assert the nature of the *frbroo:F3ManifestationProductType/bf:Instance* through the specific classes *bf:Electronic* or *bf:Print*.

An area which needs testing and implementation is the alignment of the FRBRoo *F1 Work* and *F2 Expression* classes, the latter of which refers to a different verbal form of the contents (e.g. a translation). The *F2 Expression* class constitutes a further layer, which cannot be directly mapped to a Bibframe class. One might argue that this layer can be discarded, but it is, however, relevant in our ontology. In fact, even though the diplomas in question are all conveyed in the same language, we find a distinction between *F1 Work* and *F2 Expression* relevant since redactions of the same diploma might present variant readings. We propose to use this layer to conceptualize the dMGH redaction of a given diploma.

Intuitively, the FRBRoo class *F5 Item* might be directly mapped to Bibframe's *Item* class, but on further reflection, we decided to assign the *F5 Item* and the *F4 Manifestation singleton* classes as subclasses of Bibframe's *Item* class. In fact, the *frbroo:F5Item* class does not overlap semantically with *bf:Item* class. The *frbroo:F5Item* class refers to a single item of a work which is produced in series (e.g. a given copy of a printed book), whereas the *bf:Item* class refers to any single copy of a work, including thereby also manuscripts, which are formalized in FRBRoo with the *F4 Manifestation Singleton* class.

Harmonizing these two models presents the advantage of keeping CIDOC-CRM's original event-centric formalization and enriching the list of classes with a standardly accepted vocabulary. Moreover, it has been shown by [13] that mapping between FRBR and Bibframe is possible and does not usually result in loss of semantic descriptive force, once properties are adequately assigned to the different classes (cf. also the FaBIO Ontology for an example of an integrated FRBR ontology, [10]). I will argue that harmonizing the FRBRoo and the Bibframe models, which has not been undertaken yet in my knowledge, could lay the foundations for a reusable model, which goes beyond the specific portion of real-world that is being described within the OntoVE project.

The harmonization of the models allows furthermore attaching different kinds of information to different layers, as exemplified in Figure 1. For instance, the Bibframe *originPlace* and *originDate* properties, relating to the context of

---

<sup>2</sup> A reviewer wonders why I did not decide to reuse the FaBIO Ontology [10]. This ontology, which constitutes a further example of an FRBRoo-aligned ontology, is not appropriate for the description of the items in our KB, since it is specifically designed for items that are the result of scholarly publication. For instance, the FaBIO *manuscript* class does not describe historical documents, but rather the manuscript version of e.g. a research paper. In our ontological module, the main aim is to separate the information regarding the conception of historical charters and the information regarding the published form included in the KB. The latter pieces of information also includes claims about different redactions of the same work, and potential interventions by compilers or transcribers of the same diploma. The FRBRoo quadripartite conceptualization of a bibliographic record is relevant for the goals pursued by the KB described, and constitutes a standardly accepted ontological module. This ontology, which is abstract enough to capture the different conceptual phases around the production of a written document, is complemented with finer-grained classes and properties from Bibframe.

production and ideation of the work, are attached to the *frbroo:F1Work/bf:Work* layer in order to give place and time coordinates for the conception of the work. The FRBRoo model foresees an *F27WorkConception* class and different properties relating to the creation of a work; however, the *originDate* and *originPlace* properties devised in the Bibframe model allow to directly assign to the diploma the date and place of issuance reported in the diploma itself. In fact, few information might be retrievable regarding the original conception of the act being issued with a given diploma. The alignment of the two ontological models allow to define contextual information more precisely.

The dates of publication of the critical editions, or the handles of their digitized versions, are instead modelled within the *frbroo:F3ManifestationProductType/bf:Instance* layer. This layer is also relevant for modelling the information conveyed in the editions regarding other redactions of the same work, as shown in Figure 1, where the property *bf:references* explicitly states that an *frbroo:F4ManifestationSingleton* individual is cited in the dMGH edition. This layer is also relevant for the encoding of statements about falsifying of (parts of) a given diploma, or claims about interpolation from different sources. For reasons of space, the properties relating the abstract work described in Figure 1 and this particular instance of *F4ManifestationSingleton* are not represented in Figure 1.

Since the single diplomas are collected in a unique volume, in our modelling we assign the description of the whole digitized volume to the *frbroo:F3ManifestationProductType/bf:Instance* layer and we model the information regarding the single pages conveying the redaction of a given diploma as items, as can be seen in Figure 1.

Finally, not only is an event-centric model relevant for the description of the events concerning the production of a document, as argued for above, but is also relevant for the description of the contents of the diplomas object of the present paper. In fact, such diplomas not only report about the bestowing or withdrawal of certain benefits, but also consist in their actuation, since the contents written are then effectively undertaken (or meant to be) in the real world. Therefore, one can view such documents as performative acts, and consequently model the act, which constitutes the content of the diploma, as an event. Each event is then linked to the actors involved. At the same time, these events are also modelled as subjects of the diploma itself, through the Bibframe *subject* property, which is also used to link the work to the places it refers to. This approach exploits both FRBRoo and Bibframe classes and properties to model the contents of a given diploma as performative events and as topics of it at the same time, thereby allowing to link the contents of a work to the places they are related to (cf. [1] for a similar approach). In this way, the ontology can be later queried via a SPARQL endpoint and the nature of the acts involving specific places can be further analysed by scholars and a wider audience interested in the relationship between the Veneto territories and the Germanic populations ruling over it. Finally, I will argue that reusing already established ontologies and harmonizing them constitutes a more economic approach in the field of LOD, rather than using *ad hoc* classes and properties, because it boosts interoperability with ontologies based on wide-spread standards.

### 3. FUTURE PERSPECTIVES

As mentioned in Section 1, this ontological module is embedded within the wider OntoVE project aiming at describing Germanic cultural heritage artifacts in the Veneto region; the KB under construction within the OntoVE project has the goal of providing a uniform shared description of such items, to make the ontologies interoperable with both national and international standards.

The module described in Section 2 will be subject to further testing through a local SPARQL endpoint, with the goal of assessing the semantic solidity of the harmonization of the reused ontologies and the adequacy of the model with respect to the competency questions underlying the KB.

As mentioned in Section 1, the approach undertaken in the module described in this paper was also adopted for the formal description of archaeological findings: in the project, formal ontologies such as CIDOC-CRM are reused and complemented with finer grained ontologies, such as ArCo, Architecture of Knowledge [5] for the archaeological module. The methodology employed for both models aims at aligning CIDOC-CRM and FRBRoo, which are abstract enough to capture a diverse array of items, with ontologies designed for a specific portion of knowledge. In this way, interoperability with both national and international standards is ensured, while at the same time a controlled vocabulary is used to define the individuals included in the KB.

I will argue that our approach can be further applied to different domains of knowledge and can serve as the basis for interactive analytical user interfaces. In fact, we aim to provide the users not only with a SPARQL endpoint, but also with a user interface where data are displayed via different perspectives, such as place or time maps, or aggregations of diplomas according to the rulers they were issued by. The goal is to let the user have a glance on the different acts performed by a given ruler, in an approach inspired by the Sampo Model for Letters [6].

The ontological module described here is carefully planned in order for it to be consistent with future implementations that can offer the users dynamic analytic tools.

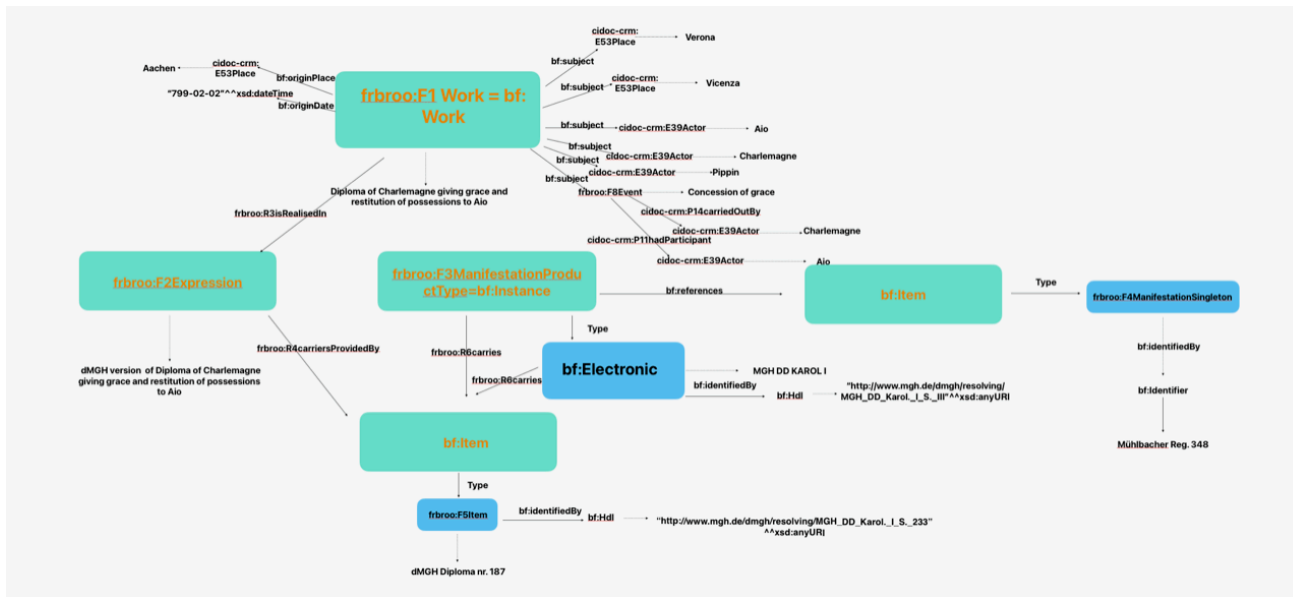


Figure 2. The OntoVE modelling exemplified through one diploma.

## 4. ACKNOWLEDGEMENTS

This project is funded by the *Programma Operativo Nazionale (PON) 2014-2020, Asse IV, Azione 4*, MUR. I thank Chiara Aiola, Alessandro Bertozzi from Net7 and Carlo Teo Pedretti for their technical assistance and the discussions on the modelling practice and on the visualization interface.

## REFERENCES

- [1] Adorni, Giovanni, Marco Maratea, Laura Pandolfo, and Luca Pulina. «An Ontology for Historical Research Documents». In: ten Cate, B., Mileo, A. (eds) *Web Reasoning and Rule Systems*. RR 2015. *Lecture Notes in Computer Science*(2015), vol 9209. Springer, Cham. [https://doi.org/10.1007/978-3-319-22002-4\\_2](https://doi.org/10.1007/978-3-319-22002-4_2)
- [2] Barzaghi Sebastian, Monica Palmirani, and Silvio Peroni. «Development of an Ontology for Modelling Medieval Manuscripts: The Case of Progetto IRNERIO». *Umanistica Digitale*, (2020): 117-140. <https://doi.org/10.6092/ISSN.2532-8816/11187>
- [3] Bekiari, Chrysoula , George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead, and Athanasios Velios. *Definition of the CIDOC Conceptual Reference Model, Version 7.1.2*, 2022.
- [4] Bekiari, Chrysoula, Martin Doerr, Patrick Le Bœuf, and Pat Riva. *FRBR, object-oriented definition and mapping from FRBRER, FRAD and FRSAD*, approved by IFLA, Version 2.4. 2015.
- [5] Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. «ArCo: The Italian Cultural Heritage Knowledge Graph», 11779:36–52, (2019). [https://doi.org/10.1007/978-3-030-30796-7\\_3](https://doi.org/10.1007/978-3-030-30796-7_3)
- [6] Hyvönen, Eero, Petri Leskinen, and Jouni Tuominen. «LetterSampo – Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data». *Journal on Computing and Cultural Heritage*, (2022), 3569372. <https://doi.org/10.1145/3569372>
- [7] Le Bœuf, Patrick. «Modeling Rare and Unique Documents: Using FRBR OO /CIDOC CRM». *Journal of Archival Organization* 10, no. 2 (April 2012): 96–106. <https://doi.org/10.1080/15332748.2012.709164>
- [8] Library of Congress, Bibframe, <https://www.loc.gov/bibframe/>
- [9] Musen, Mark A. «The Protégé project: A look back and a look forward». *AI Matters*. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), (June 2015). DOI: 10.1145/2557001.25757003
- [10] Peroni, Silvio and David Shotton. «FaBiO and CiTO: Ontologies for describing bibliographic resources and citations». In *Journal of Web Semantics* 17. Pp. 33-43. 2012. ISSN 1570-8268, <https://doi.org/10.1016/j.websem.2012.08.001>.
- [11] Possenti, Elisa. «Necropoli di Eta longobarda nel Veneto» *Quaderni Friulani di Archeologia* XI, fasc. XI (2001): 133–52.
- [12] Zapounidou, Sofia, Michalis Sfakakis, and Christos Papatheodorou. «Highlights of Library Data Models in the Era of Linked Open Data». In *Metadata and Semantics Research*, a cura di Emmanouel Garoufallou e Jane Greenberg, 390:396–407. *Communications in Computer and Information Science*. Cham: Springer International Publishing, (2013). [https://doi.org/10.1007/978-3-319-03437-9\\_38](https://doi.org/10.1007/978-3-319-03437-9_38)
- [13] Zapounidou, Sofia, Michalis Sfakakis, and Christos Papatheodorou. «Preserving Bibliographic Relationships in Mappings from FRBR to BIBFRAME 2.0». In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds) *Research and Advanced*





# Mappare il museo in IIF. Una combinazione di deep zoom e VR360 per la Galleria Borghese di Roma

Stefania De Vincentiis<sup>1</sup>, Martin Critelli<sup>2</sup>

<sup>1</sup> Università Ca' Foscari Venezia, Italia – stefania.devincentiis@unive.it

<sup>2</sup> Università Ca' Foscari Venezia, Italia – martin.critelli@unive.it

## ABSTRACT

L'adozione del framework IIF, International Image Interoperability Framework, presenta all'interno dello scenario italiano una linea discontinua e ancora frammentata, una disomogeneità dovuta soprattutto alla fase ancora iniziale nel processo di digitalizzazione delle collezioni. Tuttavia, l'enfasi su come descrivere e interpretare le collezioni storico-artistiche avvalendosi degli applicativi IIF catalizza in maniera crescente l'attenzione di musei, specie nei casi in cui questi si presentano come palazzi storici, monumenti essi stessi oltre che contenitori di arte, fortemente connotati nei loro spazi da decori, stucchi, affreschi e arredi. Il progetto in esame riguarda proprio uno di questi istituti, nella fattispecie la Galleria Borghese di Roma, e coinvolge ricercatori e studenti del Venice Centre for Digital and Public Humanities di Venezia (VeDPH), nella descrizione in IIF delle sale e delle opere digitalizzate del museo, attingendo al catalogo online delle collezioni. Le ricadute di questa sperimentazione possono interessare sia un pubblico generalizzato che specialistico, rivolto soprattutto a professionisti museali. Per questi ultimi l'adozione di tale modello può intervenire nella progettazione di mostre temporanee e nello sviluppo di attività educative avvalendosi di ulteriori applicazioni di editing adottate da altre piattaforme dello stesso framework quale lo strumento Storiies [11] [3].

## PAROLE CHIAVE

IIF; knowledge organization, digital museum, Galleria Borghese, descriptive metadata

## 1. INTRODUZIONE

Lo scenario sul IIF per l'ambiente museale è al momento in una fase di avvio<sup>1</sup>, dove molte istituzioni approfittano degli applicativi esclusivamente per rendere accattivante e dettagliata l'immagine del singolo oggetto d'arte, nell'intento di stimolare l'interazione con l'utente remoto. Questo atteggiamento si proietta in un approccio molto basilare agli strumenti IIF denotando una trascurata attenzione a nuove implementazioni e alle occasioni di confronto aperte dalla Museum Community IIF. Un dato che si evince già a partire da una rapida disamina dei diversi cataloghi online dei musei che rileva una non sempre univoca intenzione a esporre il Manifest dell'oggetto digitale, vero e proprio documento che declina nel web le informazioni disponibili relative a immagini e dati dell'opera d'arte. La creazione di una lista aggiornata e organizzata delle risorse digitalizzate comprensive di Manifest è uno dei tasselli utili a garantire l'efficacia di questo metodo come propulsore della ricerca a partire dalle raccolte digitalizzate.

Tra i musei che rendono più agevole la consultazione dei loro Manifest si annoverano il Getty Museum, l'Harvard Art Museum, il Cooper-Hewitt Smithsonian Design Museum, per citare i più completi. Una attitudine questa che rende difficile lo studio dei metodi di interazione adottati dalla singola istituzione per descrivere le opere e codificare in maniera univoca l'oggetto digitale.

Al momento il numero di istituzioni che hanno adottato tale modello, con usi tra loro differenti in base al tipo di politica sulla gestione delle immagini per le collezioni adottata dal museo, è in crescente aumento e annovera, tra gli altri, l'Art Gallery di Ontario, l'Art Institute di Chicago, il Carnegie Museum of Art, l'Hill Museum and Manuscript Library (HMML), la National Gallery of Art, il Paul Mellon Centre, la Yale Center for British Art, la Frick Collection, oltre ai già menzionati J. Paul Getty Trust, l'Harvard Art Museums, il Cooper-Hewitt Smithsonian Design Museum [1]. Sul territorio nazionale, le due istituzioni che hanno adottato il IIF per le loro raccolte sono i Musei Vaticani e le Gallerie Estensi. Entrambe hanno avviato l'opera di rimodulazione dei metadati a partire dai propri fondi archivistici e documentali: per la Biblioteca Apostolica Vaticana il progetto delle Digital Vatican Library [23] e per la Biblioteca Estense l'Estense Digital Library [8]. Rispetto alla funzione prettamente descrittiva proposta dall'adozione del IIF, la Biblioteca Estense integra lo strumento Storiies, adattato nella funzionalità Storie dalla Media Library Online-MLOL [6]. Si tratta di un applicativo che adopera

---

<sup>1</sup> Salarelli evidenzia correttamente come il IIF si sviluppi nel solco degli studi di biblioteconomia, già preparati in termini di standard procedurali comuni e condivisi, basti pensare a esempi di progetti come la DPLA o Europeana. Salarelli 2017 [21] p.3.

sequenze di immagini in IIIF per creare percorsi virtuali all'interno di una stessa immagine, selezionandone i dettagli o accostando altri da immagini differenti. I repertori a cui attingere per creare la singola Storia sono quelli della piattaforma MLOL, oppure quelli in IIIF presenti in altri progetti che condividono lo stesso framework IIIF, con l'opzione di intervenire e personalizzare con annotazioni e commenti ogni immagine incorporata.

La sperimentazione, tuttora in corso, presso il Venice Centre for Digital and Public Humanities dell'Università Ca' Foscari di Venezia segue un simile approccio narrativo. Si distingue, però, nel presentare uno nuovo sforzo di interpretare le potenzialità del framework adottando come caso studio la Galleria Borghese, considerandone la componente collezionistica e spaziale<sup>2</sup>. Infatti, oltre al patrimonio monumentale, il museo detiene un duplice carattere distintivo, caratterizzato da un ricco apparato decorativo che fa da cornice alla collezione di sculture, dipinti e arredi. Tale conformazione rende la sua descrizione in IIIF sfidante perché consente di comprendere non solo il singolo oggetto digitale ma anche la componente spaziale.

Questo aspetto è suggerito da un contesto operativo che interessa il museo e la sua resa virtuale. A partire dal 2020, infatti, la Galleria Borghese, ha avviato una campagna di digitalizzazione che ha interessato, e interessa tuttora, la componente strutturale dell'edificio e delle sue collezioni. Il catalogo digitale del museo ha seguito l'aggiornamento delle informazioni di ciascuna categoria di opere d'arte e la rispettiva organizzazione in un modello di database relazionale. In parallelo si è svolto il rilievo architettonico dell'edificio e dei suoi ambienti per la realizzazione di un modello H-BIM Historical o Heritage Building Information Modelling<sup>3</sup>. Obiettivo di questa operazione è, nel complesso, quello di concepire un modello federato per la gestione di eterogenei repertori di informazioni utile a realizzare una rappresentazione virtuale degli spazi museali e delle sue collezioni mediante l'integrazione di differenti modelli e sistemi di acquisizione di dati. Uno degli esiti del lavoro di rilievo è la ricostruzione tridimensionale degli spazi museali accessibile online nel formato di un tour virtuale. Al fine di restituire una percezione realistica delle sale, il modello integra degli scorci fotografici equirettangolari, suggerendo dei punti di vista controllati e le foto sferiche delle principali opere d'arte. Tale accorgimento presenta ulteriori risvolti: a livello metodologico, lo spazio mette a confronto una ricostruzione grafica in 3D, ottenuta dai dati di rilievo tramite laser scanner, con una composizione di immagini fotografiche equirettangolari formata da molteplici riprese e secondo diversi punti di osservazione. A livello interpretativo e della fruizione utente, il modello offre due accostamenti di realtà, una simulata, in 3D e una realistica, fotografica. La prima è caratterizzata da un massimo movimento del visitatore nello spazio, a sei gradi di libertà (6DoF), propria della Realtà Virtuale; la seconda, specifica dei video 360°, compensa i minori gradi di libertà, (3DoF) con la più alta definizione delle immagini acquisite che permette un avvicinamento all'opera e un fedele dettaglio di ogni componente stilistica e materica<sup>4</sup>. Gli obiettivi di tale modello superano, quindi, la sola rappresentazione del museo all'interno di tour virtuale. Entrambe le rese spaziali, 3D e 360°, potranno infatti essere gestite attraverso una piattaforma H-BIM, collegando le sale ai database delle collezioni consentendo, direttamente dall'ambiente virtuale, un accesso alla conoscenza delle opere esposte e alle informazioni di quelle conservate nei depositi.

La presenza di riprese fotografiche a 360° degli ambienti è la condizione che ha stimolato la sperimentazione in corso, aprendo a nuove possibilità nell'utilizzo del framework IIIF, per finalità narrative, museografiche e curatoriali. L'espedito adottato per la sperimentazione consiste nell'avvalersi di tali fotografie per la rappresentazione di ogni sala, descritta in IIIF e interrogata come una mappa, quale singolo oggetto digitale. In questa veste, le immagini si prestano ad essere indagate e commentate, guidando l'utente in una modalità amplificata del tour virtuale.

I primi risultati di questo processo per la descrizione di immagini equirettangolari presenteranno un confronto di due versioni: la prima che utilizza l'ultimo aggiornamento dell'API Presentation in 3.0 e una suddivisione per collezioni; la seconda che adotta il Simple Annotation Server. In generale, la struttura API di presentazione IIIF consente ai creatori di costruire relazioni diverse tra le risorse.

## 2. OBIETTIVI

In questo caso di studio l'obiettivo perseguito è quello di integrare le molteplici relazioni che intercorrono all'interno dell'intero catalogo di opere della Galleria Borghese nei differenti Manifest del IIIF. Tali modelli sono stati definiti come un percorso interconnesso, strutturato in relazioni tutto-parte e basato sulla reale articolazione della collezione per garantire agli utenti una esperienza di navigazione online altamente immersiva. Questo approccio sarà di supporto agli utenti nella scoperta del museo permettendo loro di muoversi virtualmente, da remoto, tra le sale e le opere d'arte in esse contenute.

<sup>2</sup> Alcuni aspetti del caso studio in esame in relazione ad altri casi museali sono stati trattati in: De Vincentis 2023 [13], p.256.

<sup>3</sup> La realizzazione del modello H-BIM è a cura del dipartimento di Architettura dell'Università la Sapienza di Roma. Per una disamina di risultati si veda: Bianchini, Inglese, Ippolito, Murrari, 2023 [7].

<sup>4</sup> Un approccio simile per interrogare l'opera d'arte attraverso riprese panoramiche e dispositivi per la Realtà Aumentata è stato affrontato in De Vincentis, Vascon, 2020 [12].

### 3. METODOLOGIA

La realizzazione del modello IIIF per la collezione della Galleria Borghese si sviluppa all'interno di un ambiente altamente interoperabile che, per sua struttura, definisce un approccio libero alla metadattazione. Così facendo, si consente la creazione di un'unica struttura valida per i differenti domini culturali a discapito, però, di una descrizione uniforme e standardizzata degli oggetti digitali [16]. Non manca tuttavia, sotto il profilo semantico, la possibilità di accedere a descrizioni standardizzate esterne, tramite la proprietà *seeAlso* che consente di creare all'interno del modello IIIF un collegamento a risorse quali schede descrittive e ontologie di riferimento. Il solo ricorso a questo elemento, nonostante consenta di ridurre il rischio di perdita di informazioni, non è sufficiente a garantire una descrizione coerente delle risorse. Per tale ragione, in questo progetto si è proceduto verso la definizione di un modello di Manifest uniforme, sotto il profilo descrittivo, allo standard di metadati Visual Resources Association Core (VRA Core) [17]. L'adozione di questo standard costituisce un punto di forte innovazione all'interno della stessa comunità IIIF laddove, nonostante la forte centralità dell'interoperabilità, *"It explicitly does not aim to provide metadata that would allow a search engine to index digital objects"* [5]. L'applicazione dello standard VRA Core rappresenta un primo step verso la definizione all'interno dello stesso modello IIIF di chiavi di accesso normalizzate, uniformi e specifiche del dominio museale e culturale in genere. Per procedere alla realizzazione di tale modello, il flusso di lavoro è stato organizzato in tre macro-fasi, così suddivise:

1. Analisi della struttura IIIF e confronto con la collezione della Galleria Borghese;
2. Definizione degli strumenti da adottare;
3. Implementazione del modello di Collection IIIF della Galleria Borghese.

#### **Fase 1. Analisi della struttura IIIF e confronto con la collezione della Galleria Borghese**

Il punto preliminare e propedeutico per la realizzazione del progetto ha riguardato l'analisi delle caratteristiche strutturali e tecniche delle API dell'IIIF. In questa fase iniziale l'attenzione è stata rivolta in particolare verso le core API di Image e Presentation, attualmente aggiornate alla versione 3.0. Rispetto alla precedente e ampiamente utilizzata, la versione attuale delle Presentation API si differenzia perlopiù sotto il profilo sintattico-strutturale e per la possibilità di includere al proprio interno documenti audio-video che, sino a quel momento, era possibile collegare solamente come risorse esterne [4]. Un ulteriore elemento valutativo emerso in questa fase, come verrà declinato specificatamente in quella successiva, ha riguardato la compatibilità tra le nuove versioni delle API e gli strumenti per l'implementazione del modello IIIF. La scelta del modello più opportuno da adottare è stata condotta sulla base della struttura dei dati all'interno delle collezioni digitali della Galleria Borghese. Si è proceduto dunque a indagare il database del museo, analizzandone le diverse classi di oggetti culturali e i metadati ad essi associati, concentrando l'attenzione verso le informazioni descrittive in chiave di uniformità e normalizzazione all'interno del dominio. Contestualmente, trattandosi di uno standard specifico per la descrizione di oggetti culturali visivi e delle immagini associate, sono stati esaminati i metadati del VRA Core, individuando tra questi un application profile da associare alla collezione del museo che risultasse congruo e idoneo alla struttura del modello IIIF. In seguito, è stato portato avanti un confronto e benchmarking con le attuali implementazioni IIIF nei diversi settori culturali, focalizzando l'attenzione al contesto museale. Da questa attività è emerso uno scenario complesso e variegato laddove, ad un progressivo sviluppo di applicazioni e strumenti, non ha fatto seguito da parte delle istituzioni culturali l'aggiornamento dei propri modelli IIIF secondo le attuali specifiche. Difatti, se nel dominio bibliotecario un gran numero di istituzioni hanno adottato le nuove versioni (Bodleian Library [9], Stanford [21] etc.), diverso è il caso dei musei dove, fatto salvo poche eccezioni, si è preferito restare ancorate alla precedente versione 2.x con la conseguente impossibilità di sfruttare a pieno il potenziale che l'ambiente IIIF offre.

#### **Fase 2. Definizione degli strumenti da adottare**

Nella fase successiva all'analisi dello stato attuale e delle infrastrutture, la ricerca è stata condotta verso la definizione degli strumenti da adottare per la realizzazione del modello IIIF. A tale scopo, è stato necessario selezionare un image server che soddisfacesse i requisiti dell'infrastruttura IIIF per la realizzazione e gestione delle immagini attraverso l'Image API. L'image server attribuisce all'oggetto immagine un identificativo (URI) a cui è possibile accedere tramite una richiesta con protocollo HTTP(S). L'applicativo Image consente l'accesso alla risorsa digitale secondo precisi parametri ordinati in: *region, size, rotation, quality e format*. La richiesta di accesso deve essere dunque formulata selezionando tali parametri sino a comporre un URI così strutturato: [http://www.server-name.org/iiif/2/image\\_nome.png/full/full/90/bitonal.jpg](http://www.server-name.org/iiif/2/image_nome.png/full/full/90/bitonal.jpg). Nell'esempio riportato, il risultato restituito sarà l'intera immagine con estensione ".jpg", mantenendo le dimensioni originali, ruotata in senso orario di 90° e nel formato bitonale. Come specificato nella stringa, l'accesso viene effettuato impiegando l'API 2.0.

Per il caso della Galleria Borghese, la selezione dell'immagine server più adatto è stata compiuta analizzando gli strumenti compatibili e presentati all'interno della stessa community IIF [2] scegliendo, tra questi e a seguito di una valutazione costi/benefici, il Cantaloupe Image Sever [10] che permette la realizzazione e la messa a disposizione di immagini in alta definizione sia attraverso la versione 2.x che 3.0 dell'Image API.

Selezionato l'immagine server, nella fase successiva ha avuto inizio la costruzione della struttura cardine del modello IIF, la Presentation API, partendo dalla selezione della versione da impiegare. Ancora una volta, indispensabile è stata una preliminare operazione di benchmarking sulle diverse implementazioni del IIF nel dominio museale, selezionando casi d'uso realizzati con le versioni 2.x e 3.0 dell'applicativo. Tra questi, è stata prestata particolare attenzione al Getty Museum [15], per la versione 2.0, e al National History Museum di Londra [19] per quella aggiornata. Alla base di questo processo valutativo sono stati tenuti in considerazione alcuni fattori quali: la congruità della versione con la collezione in essere, la compatibilità con gli strumenti esistenti, il grado di innovazione nel contesto di applicazione. A seguito delle valutazioni condotte, si è scelto di costruire il modello applicando la versione 3.0 delle Presentation API facendo sì che il progetto si configurasse tra i primi, nell'ambito museale italiano, ad adottare questa versione.

Sotto il profilo tecnico, la visualizzazione del modello IIF richiede il ricorso ad appositi visualizzatori. Questi strumenti permettono la visualizzazione del modello IIF, una volta importato in formato JSON, secondo diverse e ben specifiche caratteristiche. Attualmente, lo scenario presenta un'ampia disponibilità di visualizzatori, sia proprietari sia , integralmente o parzialmente, open source. Ai fini del presente progetto, si è preferito optare per un applicativo open source scegliendo, dopo un'attenta valutazione su funzionalità e prestazioni nell'ambito museale, il viewer Mirador 3 [18]. La disponibilità in open source e il diffuso impiego, unitamente all'elevata inclusività della community nello sviluppo e aggiornamento del software, rendono Mirador tra i più completi visualizzatori nel contesto dell'IIF. Un fattore aggiuntivo che ha spinto verso la scelta di tale viewer è la possibilità di essere personalizzato in locale attraverso la messa a disposizione di diversi plugin sempre aggiornati.

### Fase 3. Implementazione del modello di Collection IIF della Galleria Borghese.

All'interno di questo caso applicativo, il ruolo centrale è ricoperto dalla struttura della Presentation API versione 3.0. Per la rappresentazione della collezione della Galleria Borghese è stato adottato un approccio bottom-up con l'obiettivo di trasporre in IIF l'attuale suddivisione in sale del museo. In tal senso, si è proceduto a definire la struttura di Manifest, perno centrale dell'API attraverso cui un oggetto viene descritto e rappresentato, delle singole opere della collezione. In particolare, lo scopo di questa fase ha riguardato la realizzazione di un modello facilmente accessibile e interrogabile dall'intera comunità di studiosi e utenti di ogni genere, che rispondesse ai requisiti di interoperabilità del IIF e, al contempo, alle esigenze di uniformità e standardizzazione descrittiva all'interno del contesto museale. Per raggiungere tale scopo, si è definito il modello sotto il profilo descrittivo attraverso lo standard di dominio VRA Core, la cui distinzione all'interno dei tre elementi Collection, Work e Image consente di stabilire punti di connessione con il modello IIF.

```
"metadata": {
  {
    "titleSet": {
      "title": {
        "Paolina Borghese Bonaparte come Venere vincitrice"
      }
    },
  },
  {
    "dateSet": {
      "date": {
        "creation"
      },
      "earliestDate":["700"],
      "latestDate":["800"]
    }
  }
}
```

Figura 1. Esempio di metadattazione a livello di Work per il Manifest della scultura "Paolina Borghese Bonaparte come Venere Vincitrice".

Per creare dinamicità all'interno del modello rendendo la struttura maggiormente interoperabile e, di conseguenza, agevolare la navigazione agli utenti, sono stati creati dei collegamenti ipertestuali tra i Manifest delle singole opere. Al fine di valorizzare questo aspetto all'interno del progetto, si è partito dalla definizione dei Manifest delle sale del museo attraverso immagini a 360° delle stesse. Il ricorso a queste immagini, unitamente ai collegamenti ipertestuali ai Manifest, ha costituito il punto iniziale di accesso alle singole opere contenute nelle sale e che trova, nella definizione dell'elemento Collection, la sua massima espressione. Collection è l'elemento di più alto livello delle Presentation API all'interno del quale si raccolgono i Manifest consentendo, nella fattispecie del presente caso studio, di visualizzare le opere della sala nel

loro insieme, mantenendo la reale suddivisione del museo. La definizione del Collection è stata elaborata per consentire nella successiva in fase di visualizzazione, indipendentemente dal viewer impiegato dall'utente, la navigazione tra le diverse opere partendo dalla sala di riferimento. In tal modo si suggerisce un tipo di navigazione tra le sale simile a quella di un tour virtuale. (Figura 2).

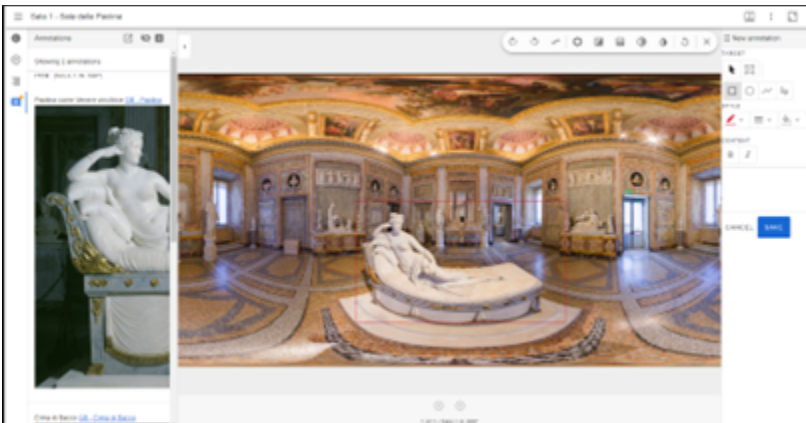


Figura 2. Esempio di visualizzazione del modello con il software Mirador 3.

Conclusa la fase di definizione dei modelli di Manifest e di Collection, si è passato alla personalizzazione dell'applicativo Mirador 3 con l'obiettivo di rendere accessibile il materiale creato a partire dal sito internet della Galleria Borghese. Questa iniziativa risponde alle esigenze di studiosi di dominio e dei singoli utenti in generale, favorendo le attività di recupero, ricerca e analisi sulle risorse. Risponde a tali necessità la nuova, implementata versione, di Mirador 3, alla quale sono stati applicati specifici plugin che rendono possibile all'utente di creare annotazioni personali, anche sotto il profilo semantico, all'interno dell'immagine stessa. L'inserimento di questi specifici plugin è stato pensato per rendere le annotazioni salvabili localmente o, tramite un apposito storage server, in remoto e visibili ad altri utenti, promuovendo così un ambiente altamente collaborativo e interoperabile.

#### 4. RISULTATI E PROGETTI FUTURI

Il progetto è stato portato avanti seguendo le buone pratiche delle Digital Humanities [22] per rendere disponibili *tout court*, e non solo accessibili, gli oggetti culturali della Galleria Borghese. Il servirsi di specifici strumenti, come descritto nella fase implementativa, è stato valutato con il fine ultimo di rendere semplici gli oggetti digitali, accessibili e consultabili, rispondenti ai bisogni informativi degli utenti. Per una analisi più dettagliata, occorre contestualizzare i risultati ottenuti alla luce dell'attuale stato di perfezionamento degli strumenti compatibili con le nuove API del IIIF che, non essendo ancora giunto a compimento, comporta alcune limitazioni nell'impiego di alcune funzionalità. Tuttavia, potendo contare su una stretta e proficua collaborazione tra fruitori e sviluppatori della comunità IIIF, è stato possibile basare il progetto sulle nuove API che hanno permesso, unitamente allo standard VRA Core, di promuovere il caso della Galleria Borghese come un punto di riferimento all'interno della stessa comunità IIIF. Per lo stesso istituto museale i risultati sono stati notevoli con la possibilità di arricchire un unico ambiente, in cui il vasto patrimonio digitalizzato viene reso facilmente accessibile, usufruibile e trasversalmente consultabile. Infatti, il modello IIIF descritto favorisce diverse modalità di visualizzazione e manipolazione dei contenuti, migliorando sensibilmente l'esperienza dell'utente a vari gradi di intervento e di approfondimento. Similmente alle linee seguite da istituzioni quali l'Estense Digital Library con il già citato strumento Storie, la Vlaamse Kunstcollectie (VKC) con lo strumento di annotazioni per i condition report [24] o il Victoria & Albert Museum con il Digirati Manifest Editor [14], anche la sperimentazione in corso adotterà una strategia di crowdsourcing, avvalendosi dello strumento di annotazioni all'interno del Mirador viewer. In tal modo si faciliterà la collaborazione di studiosi da ogni parte del mondo, consentendo anche a un pubblico di studenti e appassionati di intervenire, suggerendo nuovi contenuti opportunamente filtrati e vagliati dai funzionari del museo.

La fase di avanzamento del progetto al momento prevede o la costruzione di un procedimento automatico per la creazione dei modelli di Manifest e Collection mediante il ricorso a tecniche di Machine e Deep Learning. Sarà così possibile applicare lo stesso modello alle venti sale della Galleria Borghese, fino a comprendere, in futuro, anche gli ambienti dei depositi, aprendo a nuovi accessi e a ulteriori narrazioni.

#### BIBLIOGRAFIA

[1] (IIIF) IIIF Museums community group. <https://iiif.io/community/groups/museums/>

- [2] (IIIF) International Image Interoperability Framework. Awesome International Image Interoperability Framework (IIIF). <https://github.com/IIIF/awesome-iiif#image-viewers>
- [3] (IIIF). Events. <https://iiif.io/event/2019/goettingen/program/50/>
- [4] (IIIF). Presentation 3.0 Change Log. <https://iiif.io/api/presentation/3.0/change-log/>
- [5] (IIIF). Presentation API 3.0. <https://iiif.io/api/presentation/3.0/>
- [6] (MLOL). Stories. <https://iiif.medialibrary.it/stories/>
- [7] Bianchini Carlo, Inglese Carlo, Ippolito Alfonso, Murrari Agnese, Il rilievo integrato complesso di Galleria Borghese, Roma, Gangemi 2023
- [8] Biblioteca Estense Universitaria. Estense Digital Library. <https://edl.beniculturali.it/>
- [9] Bodleian Libraries. Digital Bodleian. <https://digital.bodleian.ox.ac.uk/>
- [10] Cantaloupe. Cantaloupe Image Server. <https://cantaloupe-project.github.io/>
- [11] Cogapp. Storiies. <https://storiies.cogapp.com/>
- [12] De Vincentis Stefania, Vascon Luca. Digital languages for Art History. Audience Engagement, Virtual and Augmented Reality, in Kathryn Brown The Routledge Companion to Digital Humanities and Art History. London: Routledge, 2020: 275-286.
- [13] De Vincentis, Stefania. «Paragoni digitali per la storia dell'arte. Potenzialità e sviluppi del IIIF per le collezioni museali». Rivista d'Arte, (2023): 243-259. In corso di pubblicazione.
- [14] Fraser, Stephen, Tom Crane, Kelsie Caldwell, and Jules Schoonman. Digirati Manifest Editor. <https://github.com/digirati-co-uk/iiif-manifest-editor>
- [15] Getty. Getty Museum Collection. <https://www.getty.edu/art/collection/>
- [16] Hyam, Roger, and Quentin Groom. «Implementation of the IIIF for Natural History Collections.» 2021.
- [17] Library of Congress. «Visual Resources Association Core» 2014.
- [18] Mirador. Project Mirador. <https://github.com/ProjectMirador>
- [19] National History Museum. Museum Collection. <https://www.nhm.ac.uk/discover/collections.html>
- [20] Salarelli, Alberto. «International Image Interoperability Framework (IIIF): A panoramic». J LIS I, no. VIII (2017): 50-66.
- [21] Stanford Libraries. Digital collections. [https://searchworks.stanford.edu/catalog?f%5Bcollection\\_type%5D%5B%5D=Digital+Collection](https://searchworks.stanford.edu/catalog?f%5Bcollection_type%5D%5B%5D=Digital+Collection)
- [22] Tomasi, Francesca. Organizzare la conoscenza: Digital Humanities e Web semantico. Milano: Editrice Bibliografica, 2022.
- [23] Vatican Library. DigiVatLib. <https://digi.vatlib.it/>
- [24] Vlaamse Kunstcollectie. Project IIIF 2 WP4. <https://vlaamsekunstcollectie.be/project-iiif-2-wp-4>

# H2IOSC: Humanities and Heritage Open Science Cloud

Emiliano Degl'Innocenti<sup>1</sup>, Monica Monachini<sup>2</sup>, Alberto Bucciero<sup>3</sup>,  
Enrico Pasini<sup>4</sup>, Bruno Fanini<sup>5</sup> and Francesca Frontini<sup>6</sup>

<sup>1</sup> Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia – emiliano.deglinnocenti@cnr.it

<sup>2</sup> Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILC), Italia – monica.monachini@ilc.cnr.it

<sup>3</sup> Istituto di Scienze del Patrimonio Culturale (CNR-ISPC), Italia, Italia – alberto.bucciero@cnr.it

<sup>4</sup> Università di Torino / Istituto del Lessico Intellettuale europeo e Storia delle Idee (CNR-ILIESI), Italia – enrico.pasini@unito.it

<sup>5</sup> Istituto di Scienze del Patrimonio Culturale (CNR-ISPC), Italia – bruno.fanini@unito.it

<sup>6</sup> Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILC), Italia – francescafrontini@gmail.com

## ABSTRACT

Questo poster descrive gli obiettivi del progetto H2IOSC, Humanities and Heritage Open Science Cloud, che mira a costruire un cluster federato e inclusivo di IR nel dominio ESFRI dell'innovazione sociale e culturale volto a supportare ai ricercatori nelle varie discipline nei settori delle scienze umane, delle tecnologie linguistiche e dei beni culturali fornendo dati, strumenti e servizi digitali avanzati.

## PAROLE CHIAVE

Infrastrutture digitali di ricerca; innovazione sociale e culturale; patrimonio culturale; scienza aperta

## 1. IL CONTESTO NAZIONALE ED EUROPEO

L'Europa gode di un ricco panorama di infrastrutture di ricerca disciplinari sia nazionali che transnazionali<sup>1</sup>. A livello europeo molte di esse si raccolgono sotto il cappello dell'ESFRI (European Strategy Forum on Research Infrastructures<sup>2</sup>), la cui missione è di sostenere un approccio coerente e strategico all'elaborazione delle politiche sulle infrastrutture di ricerca in Europa e facilitare le iniziative multilaterali che portano a un migliore utilizzo e sviluppo delle infrastrutture di ricerca, a livello dell'UE e internazionale. L'Italia aderisce a diverse infrastrutture ESFRI nel settore dell'innovazione sociale e culturale, e identifica quelle prioritarie nel PNIR, Piano Nazionale delle Infrastrutture di Ricerca<sup>3</sup>, sostenendole inoltre con il pagamento delle quote di adesione e con finanziamenti per i nodi nazionali.

Il Piano Nazionale di Ripresa e Resilienza (PNRR), ha identificato le infrastrutture di ricerca come un obiettivo fondamentale da sostenere con bandi dedicati. Tra i beneficiari di questi finanziamenti vi è il progetto H2IOSC<sup>4</sup> che federa i nodi nazionali di CLARIN, DARIAH, E-RIHS e OPERAS, quattro infrastrutture ESFRI dedicate rispettivamente alle risorse e tecnologie del linguaggio, alle digital humanities, al patrimonio culturale e alla scienza aperta. Il progetto, lanciato nel novembre 2022, coinvolge 12 istituti in diverse sedi del CNR (ente responsabile per la gestione di tutte e quattro le infrastrutture); il coordinamento è affidato a CNR-OVI.

H2IOSC è attualmente nelle sue fasi iniziali, che si concentrano principalmente sulle acquisizioni di personale e strumentazione. Nei prossimi mesi il progetto metterà in campo una serie di iniziative volte a identificare le risorse e tecnologie abilitanti, che necessitano del supporto della comunità. Con questo contributo si intende quindi presentare il progetto alla comunità dell'AIUCD, che rappresenta uno dei punti di riferimento scientifici e disciplinari più importanti.

## 2. OBIETTIVO DEL PROGETTO E PROSPETTIVE

H2IOSC si inserisce nell'attuale contesto europeo che vede una sempre più stretta collaborazione delle infrastrutture disciplinari con l'EOSC (European Open Science Cloud)<sup>5</sup> che si è concretizzata nella costituzione di 5 cluster disciplinari<sup>6</sup>, tra cui SSHOC (Social Sciences and Humanities Cluster)<sup>7</sup>.

Sul piano nazionale, H2IOSC si allinea a quanto già implementato in altri paesi quali Francia, Olanda e Germania, dove infrastrutture nazionali disciplinari come Huma-Num, CLARIAH-NL o TextPlus federano e gestiscono le attività nazionali di due o più infrastrutture ESFRI, fornendo ai ricercatori risorse e servizi dedicati e sostenibili. A tal fine, H2IOSC

<sup>1</sup> Per una introduzione al panorama italiano delle infrastrutture del settore si veda [1].

<sup>2</sup> <https://www.esfri.eu/>

<sup>3</sup> <https://www.mur.gov.it/it/news/mercoledì-20102021/pubblicato-il-piano-nazionale-infrastrutture-di-ricerca-2021-2027>

<sup>4</sup> <https://www.h2iosc.cnr.it/>

<sup>5</sup> <https://eosc.eu/>

<sup>6</sup> <https://eosc-portal.eu/esfri-thematic-cluster-projects>

<sup>7</sup> <https://www.sshopencloud.eu/>



permetterà di ottimizzare l'utilizzo dei servizi nazionali e internazionali più rilevanti offerti potenziandoli e implementandoli, in considerazione dei bisogni derivanti direttamente dalle comunità scientifiche di riferimento. Permetterà in particolare di dare un'accelerazione decisiva al processo di transizione digitale nel settore delle *humanities*, favorendo l'utilizzo di strumenti e tecnologie da parte di comunità di ricercatori (tradizionalmente meno inclini al loro utilizzo) per la creazione e la gestione di dati digitali e per la loro apertura verso la società, e aiutando a mettere in rete, ampliare le strutture, gli impianti e la strumentazione a disposizione della comunità scientifica nei campi della ricerca umanistica e nel settore della preservazione del patrimonio culturale.

Come illustrato dalla Figura 1 - il progetto si articola in otto pacchetti (WP) di lavoro. Al cuore del progetto è il WP5, che vedrà la costituzione di un *marketplace* nazionale, che esporrà dati e servizi delle comunità nazionali di riferimento.

Questo verrà accompagnato da una serie di *data center* sul territorio nazionale che potenzieranno l'offerta di storage e risorse computazionali (WP4). A questo si affiancano le attività del WP1, che seguirà la gestione del progetto e assicurerà anche la sostenibilità di H2IOSC in un orizzonte decennale, del WP2, WP3 e WP6 che si occuperanno di identificare i bisogni della comunità e potenziare l'offerta (di dati e servizi), del WP7 che sarà dedicato ai Community Pilots, una serie di casi di studio su tematiche specifiche (tra cui alcuni dedicati alle edizioni digitali e all'epigrafia digitale, agli archivi sonori, all'utilizzo di tecnologie del Semantic Web e all'archeologia digitale - si veda la Figura 2 per la lista completa), e il WP8 per il fondamentale aspetto della formazione e condivisione di materiali didattici.



Figura 1. Lo schema delle attività del progetto.



Figura 2. Community pilots.

Attualmente sono già in corso contatti con le associazioni disciplinari, al fine di preparare un piano di lavoro per la ricognizione dei bisogni delle comunità. Saranno di particolare rilevanza per AIUCD la ricognizione dei bisogni per quanto riguarda le tecnologie del testo, la filologia digitale, la public history e l'open access. Molto importante sarà anche la collaborazione per quanto riguarda la raccolta dei bisogni di formazioni specifiche su strumenti infrastrutturali e metodi digitali, in particolare per i giovani ricercatori.

### 3. RICONOSCIMENTI

Il progetto H2IOSC è finanziato dal Piano Nazionale di Ripresa e Resilienza (PNRR) - Next Generation EU (NGEU) - Fondi Unione Europea

### BIBLIOGRAFIA

[1] Monachini, M. e Frontini, F. Infrastrutture digitali per le scienze umane e sociali. In Digital Humanities. Carocci, in pubblicazione.

# Śivadharma Database CMS.

## HTML and graph as a starting point for digital editions

Martina Dello Buono<sup>1</sup>, Francesca Tomasi<sup>2</sup>

<sup>1</sup> University of Bologna, Italy – [martina.dellobuono2@unibo.it](mailto:martina.dellobuono2@unibo.it)

<sup>2</sup> University of Bologna, Italy – [francesca.tomasi@unibo.it](mailto:francesca.tomasi@unibo.it)

### ABSTRACT

*Śivadharma Database* is a Content Management System (CMS) for creating, publishing, and updating Digital Scholarly Editions. By entering the *textus constitutus* first, scholars can annotate it via user-friendly tools to provide their editions with additional components, such as apparatus, notes, translation, parallels, and citations.

The annotation system is based on HTML and graph data structure. The *textus* corresponds to an HTML file ready to be displayed. Textual fragments annotated in HTML are the atomic unit of multiple relationships organized in a graph database, allowing the formal representation of complex and overlapped textual phenomena.

These methodological and practical approaches result in good annotation expressiveness and a minimum effort to learn technologies during the editing workflow by scholars.

### KEYWORDS

Digital Scholarly Editing; Text Encoding; HTML; Graph Database; CMS

## 1. INTRODUCTION

One of the most frequent issues in the Digital Scholarly Editing field is the best encoding for digital editions. XML/TEI<sup>1</sup> is the traditional solution [7]. Another proposal is the still-emerging RDF<sup>2</sup>, applied in the context of the Semantic Web. They both have pros and cons. XML/TEI is popular and easy to learn. However, the extent of its vocabulary undermines the choice of descriptors by editors and the interoperability of editions [16]. Furthermore, its limitations in representing complex textual phenomena, e.g., overlapping hierarchies, are known in the literature [10, 12, 15, 17].

On the other hand, RDF can solve the inherent problems of an XML-based tree structure, but it is a more expensive solution regarding learnability. In addition, visualization is not in the scope of RDF [1].

Another critical issue related to text encoding goes beyond its methodological implications. The encoding workflow often requires high technical skills and tasks, e.g., learning a specific language [11, 13, 14]. Scholars, instead, should have *user-friendly* software, tools, and services to streamline work.

Therefore, technological choices have an impact not only on the expressiveness of the annotations but also on the scholars' work itself. On the one hand, they must be able to manage textual complexity; on the other, they must facilitate the annotation workflow. Therefore, finding a solution that meets both needs is necessary.

The *Śivadharma Database* project proposes a solution to the highlighted issues by implementing a CMS based on HTML and graph data structure to create editions from scratch via an interface.

The present paper is structured as follows. Section 2 inspects the pros and cons of the currently available tools for text encoding. The following Section 3 reports the *Śivadharma* use case. Then, Section 4 focuses on the methodology to design the *Śivadharma Database* CMS. In conclusion, Section 5 analyzes the *Śivadharma Database* annotation system.

## 2. STATE OF THE ART

Scholars have few user-friendly tools for text encoding at their disposal. Some of these tools align their annotation systems to traditional XML/TEI for the sake of technology popularity and reusability, while others experiment with emerging technologies, e.g., RDF.

EVT<sup>3</sup> is undoubtedly the most widespread tool in the Digital Scholarly Editing field. Based on XML/TEI, it is not precisely an annotator but a viewer of digital editions. It inputs an XML/TEI file and returns its visualization via a specific interface.

---

<sup>1</sup> Text Encoding Initiative: <https://tei-c.org/>

<sup>2</sup> RDF: <https://www.w3.org/RDF/>

<sup>3</sup> EVT: <http://evt.labcd.unipi.it/>

Basic released functionalities can be checked in the pilot EVT project, *The Digital Vercelli Book*<sup>4</sup>, a diplomatic and diplomatic-interpretative digital edition. More advanced functionalities, such as the critical apparatus and parallel texts, are available in *Edizione Logica Avicennae* [experimental encoding]<sup>5</sup>. The main drawback of such a system is the required scholars' prior knowledge of the specific XML/TEI schema to visualize their editions.

In other cases, RDF is the framework on which annotation systems rely. For instance, KWICKWOCKWAC (KeyWords In Context, KeyWord Out of Context, KeyWord Alongside Context)<sup>6</sup> exploits the potentialities of RDFa<sup>7</sup>. The *Edizione Nazionale delle Opere di Aldo Moro* project [8] shows that KWICKWOCKWAC allows annotating entities, such as persons and places, in full-text documents. The results of the annotation workflow are RDFa documents.<sup>8</sup> It implies embedding the annotations into texts, thus undermining their readability.

The SPEEDy editor [9] allows the same annotations, e.g., persons and places. However, it differs from KWICKWOCKWAC in the technologies implied. Since it records the annotations as stand-off properties in Neo4j<sup>9</sup>, a graph database, it can elegantly handle complex and overlapped structures.

As KWICKWOCKWAC and SPEEDy, *Pundit*<sup>10</sup> provides a user-friendly interface to highlight and comment on web pages. An example of its use is available in the *Burckhardt Source*<sup>11</sup> project, a collection of letters from and to Jacob Burckhardt where annotations concern the persons, places, artworks, and bibliographies mentioned [3]. In addition, one specific version of *Pundit*, i.e., *Pundit Annotator Pro*<sup>12</sup>, exploits semantic technologies to create semantic annotations and link them to external resources, e.g., Wikipedia entities.

Such projects, i.e., KWICKWOCKWAC, SPEEDy, and *Pundit*, do not require coding skills since a graphical interface allows user-side annotation control. On the other hand, their annotation systems do not allow the creation of sophisticated edition components, such as the apparatus.

This task, instead, can be accomplished in the *Classical Text Editor*<sup>13</sup> environment. It supports advanced features such as the apparatus, notes, and parallel text creation. It also allows the XML/TEI export for interoperability [5]. However, its license, as in the case of *Pundit*, does not allow widespread use in the community. In addition, it is more focused on printed editions.

### 3. USE CASE

Translocal Identities: The Śivadharma and the Making of Regional Religious Traditions in Premodern South Asia, better known as the *Śivadharma Project*, is an ongoing ERC project. One of its objectives is to produce Digital Scholarly Editions of the Sanskrit texts of the so-called *Śivadharma corpus*. Written around the 6th and 7th centuries, they describe the duties a Śaiva devotee must observe [2]. They are very complex, starting from their rhythmic structure. They are in stanzas, each of which includes four quarters, Skr. *pādas*. They are conventionally called *pāda a*, *pāda b*, *pāda c*, and *pāda d*, and do not match with lines [6]. Moreover, specific textual features often break the boundaries of *pādas* and overlap. For instance, Fig. 1 shows that the string identified as the *lemma* in an apparatus entry does not fit in a single *pāda* but spans over two ones, thus undermining the traditional XML-based tree structure commonly used in text encoding.

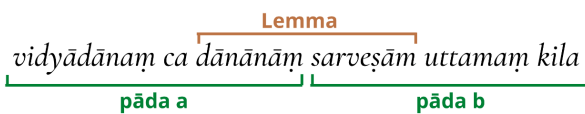


Figure 1. Lemma breaking the boundaries of *pādas*.

Other phenomena, such as omissions, lacunae, transpositions, and many translations and parallels, underline the same issue. It follows that it is necessary to design a specific data structure for text encoding to handle this complexity. Leveraging specific technologies such as HTML and a graph database, we outline methodological and practical approaches to overcome the limitations in the annotation workflow.

<sup>4</sup> The Digital Vercelli Book: <http://vbd.humnet.unipi.it/beta2/>

<sup>5</sup> Edizione Logica Avicennae [experimental encoding]: [http://evt.labcd.unipi.it/demo/evt2-beta2/avicenna/index.html#/readingTxt?d=doc\\_1&p=C-112v&s=text-body-div&e=critical](http://evt.labcd.unipi.it/demo/evt2-beta2/avicenna/index.html#/readingTxt?d=doc_1&p=C-112v&s=text-body-div&e=critical)

<sup>6</sup> KWICKWOCKWAC: <https://aldomorodigitale.unibo.it/about/docs/processing#structure-section>

<sup>7</sup> RDFa: <https://rdfa.info/>

<sup>8</sup> *Edizione Nazionale delle Opere di Aldo Moro* technical specifications: <https://doi.org/10.5281/zenodo.5144961>

<sup>9</sup> Neo4j: <https://neo4j.com/>

<sup>10</sup> Pundit: <https://www.netseven.it/pundit/>

<sup>11</sup> Burckhardt Source: <https://burckhardtsource.org/>

<sup>12</sup> Pundit Annotator Pro: <https://thepundit.it/annotator-pro-web-annotation/>

<sup>13</sup> Classical Text Editor: <https://cte.oeaw.ac.at/>

## 4. METHODOLOGY

The present research proposes HTML and graph data structure as a starting point for encoding digital editions. It is a practical strategy for several reasons. First, HTML is perfect for visualization since it does not need specific conversions. Especially if a digital edition is a short-term project, an HTML encoding can satisfy the basic request to display the edition. In addition, it can be the matrix of conversion in other languages.

The graph data structure allows the formal representation of complex and overlapping textual phenomena identified by editors. HTML can be the starting point for modelling the data of editions as a graph. Specifically, a textual fragment identified in HTML can work as a graph external database's atomic unit for multiple relations, even with external resources. So, the interpretation by editors becomes a stand-off layer, thus increasing the readability of texts and improving data interoperability [16].

Finally, such technologies can be solid pillars for building a user-friendly stand-off annotation CMS [16] based on *forms* to encode editions via an interface. A solution of this kind can allow editors to focus only on scholarly questions and decrease the potential occurrence of errors. Furthermore, it can allow complete annotation control if the text changes. The atomic unit of annotation is the fragment, which is directly selected from the text by the editor, uniquely identified, and sent to the database via a form. If the editor modifies an already identified fragment, the system detects the change by verifying its ID (Identifier) and automatically sends it back to the database. At the same time, its assigned annotation remains linked to the fragment and does not change.

A system of this kind can support the creation of even complex components of editions, e.g., apparatus, notes, parallels, citations, and translation. Since the fragment is the basic unit and it is possible to associate it with  $n$  relations, building and visualizing the apparatus, for instance, is accessible. As shown in Fig. 2, the selected fragment `:selectedFragment` is the subject of the relation `:HAS_LEMMA`. The object `:lemma` is, in turn, the subject of the relation `:HAS_VARIANT`. Then both `:lemma` and `:variant` are subjects of the relation `:ATTESTED_IN` to link them to their witnesses.<sup>14</sup>

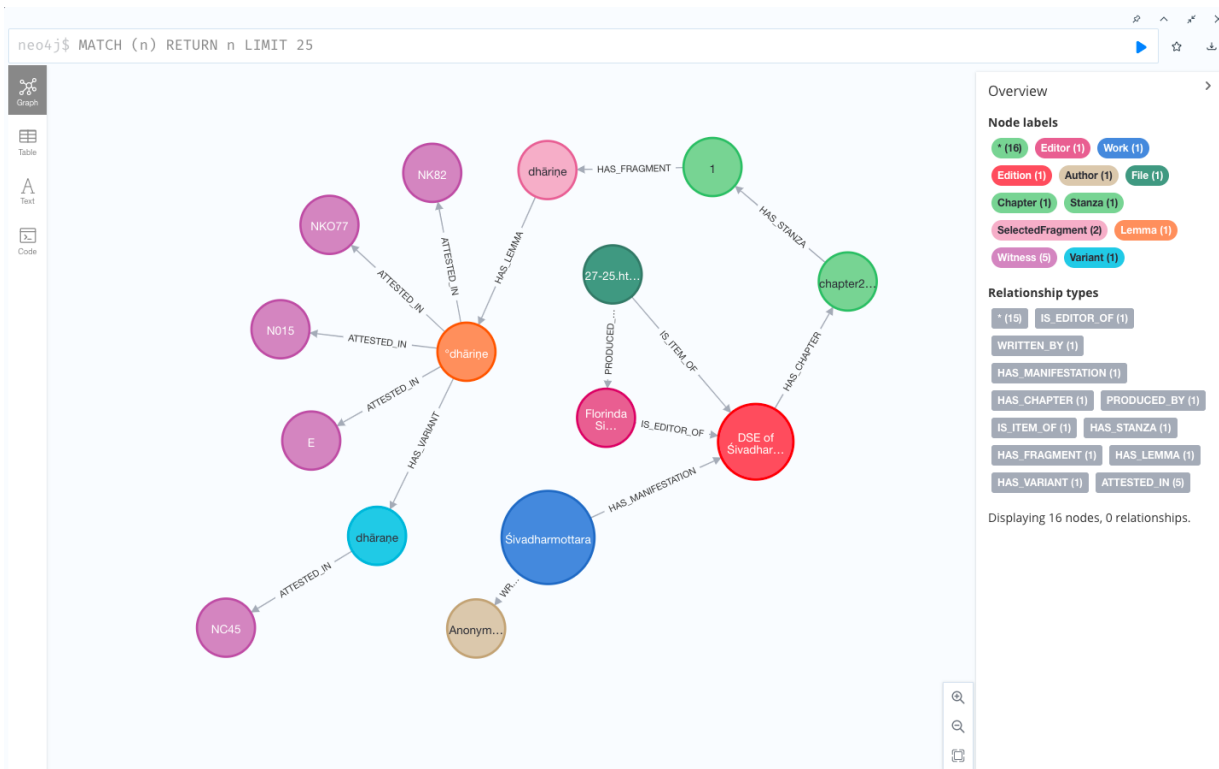


Figure 2. Apparatus entry visualized in Neo4j<sup>15</sup> graph.

Thus, we can easily visualize the apparatus entries and connect them to the text by extracting these data from the database.

<sup>14</sup> The reference language is Cypher, Neo4j's graph query language: <https://neo4j.com/developer/cypher/>

<sup>15</sup> As described in Section 5, Neo4j is the native graph database implemented in *Śivadharma Database*.

Finally, a system based on *commits* will be implemented to allow editors to track the earlier versions of their editions at each update.<sup>16</sup> Each commit will report (1) a title; (2) an agent corresponding to the actual author of the update; (3) a date; (4) an optional description of the operations performed. In this way, it is possible to trace every modification to editions and their provenance and sort them in order of time.

## 5. RESULTS

*Śivadharma Database* is a CMS to create, publish, and update Digital Scholarly Editions over time via an interface. It has been developed from scratch in Node.js<sup>17</sup>, Express<sup>18</sup>, Vanilla JS<sup>19</sup>, EJS<sup>20</sup>, SCSS<sup>21</sup>, and Neo4j<sup>22</sup>. It provides a user-friendly environment that supports the production of (1) the *textus constitutus*; (2) apparatus; (3) notes to the text; (4) translation and notes to the translation; (5) parallels, notes to the parallels, translation of the parallels, notes to the translation of the parallels; (6) citations. The data creation and management system exploits the potentiality of the *form* interface component. It is possible to send data to the graph Neo4j database in the backend by filling out specific forms for each edition component. As Fig. 3 shows, the workflow to annotate the text is basic.

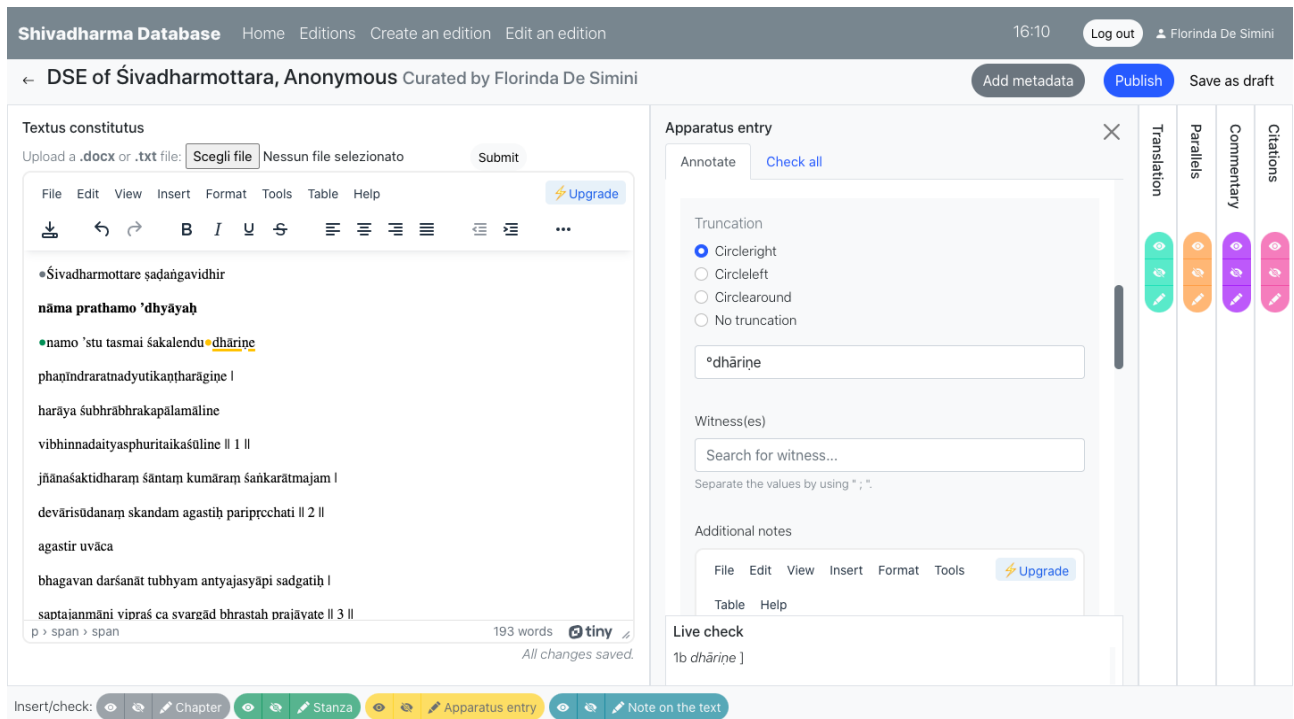


Figure 3. Forms to build an apparatus entry in *Śivadharma Database*.

The required steps are the following. First, scholars should provide the *textus constitutus* by uploading a file or typing it in the main text area. Its conversion into an HTML file and storage in a specific folder are automatic. Then, each fragment of the text can become an object of annotation. Buttons in different colors show the available types of annotations, e.g., yellow for the apparatus entries. A line in the specific color highlights the selected fragment in the text once clicked on one of them. In HTML, the selected fragment corresponds to a string surrounded by uniquely identified empty spans, which act like milestones to handle overlapping issues. Simultaneously, a module opens up on the right of the *textus*. It contains forms in which to declare specific information about the selected fragment. For instance, it is possible to create an apparatus entry by declaring the location of the selected fragment, lemma, variant readings, witnesses, additional notes, and omissions. Finally, the *Save* operation allows the submission of data to the database. Such data correspond to specific nodes

<sup>16</sup> The reference system is the GitHub system to track changes to the uploaded files: <https://docs.github.com/en/desktop/contributing-and-collaborating-using-github-desktop/making-changes-in-a-branch/committing-and-reviewing-changes-to-your-project>

<sup>17</sup> Node.js: <https://nodejs.org/en>

<sup>18</sup> Express: <https://expressjs.com/>

<sup>19</sup> Vanilla JS: <http://vanilla-js.com/>

<sup>20</sup> EJS: <https://ejs.co/>

<sup>21</sup> SCSS: <https://sass-lang.com/>

<sup>22</sup> Cfr. Note 9.

and relations automatically linked to their reference fragment, in turn, identified as a node in the database [Fig. 2]. The same system is applied for updating any data.

Regarding the visualization of editions, it exploits two different mechanisms. The *textus* corresponds to an HTML file. Thus, it automatically results in a visualization. The other components, instead, e.g., the apparatus, need intermediate data processing. First, they are extracted from the database by specific queries in the backend, sorted according to preset criteria, and finally displayed. Along with the visualization, the CMS guarantees the conversion of editions into XML/TEI to ensure their reuse and preservation.<sup>23</sup>

The *Shivadharmā Database* source code and documentation on how to run it are available on GitHub<sup>24</sup> to guarantee CMS replicability. Since it is built in modules<sup>25</sup>, the CMS customization according to the project specifics is easy. The help of a web developer is recommended to accomplish this task.

## 6. CONCLUSION

The technological choices for text encoding are pivotal since they influence the expressiveness of annotations and the effort required to learn and apply these technologies. The literature suggests two trends in Digital Scholarly Editing. On the one hand, the encoding process in the traditional XML/TEI is straightforward from the purely practical point of view of writing code. However, choosing the descriptors and handling complex textual structures are demanding tasks. On the other hand, the less consolidated RDF proposes a data structure capable of handling innate textual complexity but implies extended practical learning. In addition, the visualization does not fall within its objectives.

The present research participates in the debate proposing a solution based on HTML and graph data structure. This solution shows both methodological and practical advantages. First, HTML encoding immediately satisfies the essential requirement of Digital Scholarly Editing, i.e., the visualization of editions. In addition, it lends itself to multiple conversions into other languages, e.g., XML/TEI, for reuse and preservation purposes. On the other hand, the graph data structure guarantees a good level of expressiveness of the annotations, where the textual fragment is the pivot.

*Śivadharmā Database* implements this solution by providing scholars with user-friendly tools to create, publish, and update Scholarly Digital Editions from scratch via an annotation interface. Providing an interface reduces the scholars' required effort on purely technological issues and guarantees complete control over the annotations, even when updating the text.

Future work includes the release of additional functionalities, e.g., the automatic reconstruction of witnesses and tools for data visualization and edition filtering, and the extension of its documentation to meet reusability needs.

Since the tools available are not strictly related to the Sanskrit field, this application is valuable regardless of the application domain<sup>26</sup>. Even the replicability of the CSM itself is ensured by the online availability of its open-source code and instructions for its running.

## REFERENCES

- [1] Daquino, Marilena, Francesca Giovannetti, and 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*, no. 7 (18 December 2019). <https://doi.org/10.6092/issn.2532-8816/9091>
- [2] De Simini, Florinda. «Śivadharmā Manuscripts from Nepal and the Making of a Śaiva Corpus». In *One-Volume Libraries: Composite and Multiple-Text Manuscripts*, 233–86. De Gruyter, 2016. <https://doi.org/10.1515/9783110496956-009>
- [3] Di Donato, Francesca, and Susanne Müller. «Burckhardtsource.Org: The Unpublished Correspondence to Jacob Burckhardt». *Lexicon Philosophicum: International Journal for the History of Texts and Ideas*, no. 2 (14 March 2014). <https://doi.org/10.19283/lph-20142.409>.
- [4] Griffiths, Arlo, and Axelle Janiak. «Encoding Guide for Critical Editions. Draft Version 0.1». 2020-2023.
- [5] Hagel, Stefan. «The Classical Text Editor. An attempt to provide for both printed and digital editions». *Digital philology and medieval texts* (2007): 77–84.
- [6] Hahn, Michael. «A brief introduction into the Indian metrical system (for the use of students)». [https://www.academia.edu/6353023/Michael\\_Hahn\\_A\\_brief\\_introduction\\_into\\_the\\_Indian\\_metrical\\_system\\_for\\_the\\_use\\_of\\_students](https://www.academia.edu/6353023/Michael_Hahn_A_brief_introduction_into_the_Indian_metrical_system_for_the_use_of_students)

---

<sup>23</sup> The development of the conversion into TEI functionality is ongoing. The reference guidelines are developed within the DHARMA Project [4]: <https://dharma.hypotheses.org/a-propos>

<sup>24</sup> *Śivadharmā Database* GitHub repository: <https://github.com/martinadellobuono/shivadharmā-database>

<sup>25</sup> Specifically, a module for each kind of Scholarly Digital Edition component, e.g., one for covering the apparatus, one for translations, etc.

<sup>26</sup> The flexibility and reusability of *Shivadharmā Database* CMS will soon be tested for creating the FICLIT Digital Library collection of Italian, Latin, and Greek Scholarly Digital Editions: <https://dl.ficlit.unibo.it/s/lib/page/home>



- [7] Mancinelli, Tiziana and Elena Pierazzo. *Che cos'è un'edizione scientifica digitale*. Roma: Carocci editore, 2020.
- [8] Moro, Aldo. «Edizione Nazionale delle Opere di Aldo Moro». Voll. Università di Bologna, 2021. <https://doi.org/10.6092/unibo/aldomoro>
- [9] Neill, Iian, and Desmond Schmidt. «SPEEDy. A Practical Editor for Texts Annotated with Standoff Properties». *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing* 15 (2021): 45.
- [10] Peroni, Silvio, and Fabio Vitali. «Annotations with EARMARK for Arbitrary, Overlapping and out-of Order Markup». In *Proceedings of the 9th ACM Symposium on Document Engineering*, 171–80. DocEng '09. New York, NY, USA: Association for Computing Machinery, 2009. <https://doi.org/10.1145/1600193.1600232>
- [11] Pierazzo, Elena. «What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter». *International Journal of Digital Humanities* 1, no. 2 (1 July 2019): 209–20. <https://doi.org/10.1007/s42803-019-00019-3>
- [12] Renear, Allen, David Dubin, and C. M. Sperberg-McQueen. «Towards a Semantics for XML Markup». In *Proceedings of the 2002 ACM Symposium on Document Engineering*, 119–26. DocEng '02. New York, NY, USA: Association for Computing Machinery, 2002. <https://doi.org/10.1145/585058.585081>
- [13] Robinson, Peter. «Current Issues in Making Digital Editions of Medieval Texts—or, Do Electronic Scholarly Editions Have a Future?». *Digital Medievalist* 1, no. 0 (20 April 2005). <https://doi.org/10.16995/dm.8>
- [14] Schmidt, Desmond. «The Current State of the Digital Scholarly Edition and Three Challenges». Accessed 21 February 2023. [https://www.academia.edu/37585331/The\\_Current\\_State\\_of\\_the\\_Digital\\_Scholarly\\_Edition\\_and\\_Three\\_Challenges](https://www.academia.edu/37585331/The_Current_State_of_the_Digital_Scholarly_Edition_and_Three_Challenges).
- [15] Schmidt, Desmond. «The Inadequacy of Embedded Markup for Cultural Heritage Texts». *Digital Scholarship in the Humanities* 25, no. 3 (2010): 337–56.
- [16] Schmidt, Desmond. «Towards an Interoperable Digital Scholarly Edition». *Journal of the Text Encoding Initiative*, no. Issue 7 (12 November 2014). <https://doi.org/10.4000/jtei.979>
- [17] Sperberg-McQueen, C. M., and Claus Huitfeldt. «GODDAG: A Data Structure for Overlapping Hierarchies». In *Digital Documents: Systems and Principles*, edited by Peter King and Ethan V. Munson, 139–60. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2004. [https://doi.org/10.1007/978-3-540-39916-2\\_12](https://doi.org/10.1007/978-3-540-39916-2_12)



# L'approccio human-centered nella progettazione di un ecosistema digitale che massimizzi la diffusione del materiale diaristico preservandone la singolarità

Sara Ermini<sup>1</sup>, Martina Magri<sup>2</sup>

<sup>1</sup> Università di Siena, Italia – sara.ermini@unisi.it

<sup>2</sup> Università di Modena e Reggio Emilia, Italia – martina.magri@unimore.it

## ABSTRACT

Dal 2009 il patrimonio documentario dell'Archivio Diaristico Nazionale (ADN) di Pieve Santo Stefano è nel Codice dei Beni Culturali dello Stato e viene riconosciuto a livello nazionale e internazionale come una vera e propria "istituzione" della scrittura popolare.

In questo contesto, nasce il progetto *PDS* (Piattaforma di Digital Storytelling), con l'obiettivo di sviluppare una strategia per supportare la preservazione, la valorizzazione e la diffusione del patrimonio culturale materiale e immateriale conservato dall'ADN, nel rispetto della peculiarità della scrittura manuale e della materialità del suo supporto. Il progetto sfrutta le nuove tecnologie emergenti e le competenze culturali maturate attorno ai nuovi media [2,5,6,7] per sviluppare soluzioni e strategie in grado di massimizzare l'accessibilità, l'usabilità e il riuso creativo dei materiali conservati. Il fine ultimo è quello di incrementare la partecipazione attiva dei portatori di interesse e della cittadinanza nei processi di raccolta e divulgazione delle testimonianze autobiografiche e di favorire l'intercettazione di nuovi pubblici.

Il presente contributo si concentra sulla condivisione dell'esperienza di *PDS*, attirando l'attenzione sul processo di co-progettazione e sullo sviluppo di un protocollo replicabile - per la divulgazione digitale dei contenuti d'archivio - che ripercorre tutte le tappe del lavoro, dalla *user research* alla prototipazione della soluzione finale, includendo gli aspetti giuridici legati alla tutela della privacy e dei diritti d'autore [10].

Per progettare la piattaforma *PDS* e permettere di massimizzare la diffusione del materiale diaristico, senza normalizzare la singolarità di ogni testo e nel rispetto degli autori stessi, sono state adottate le metodologie di *human-centered design* e gli strumenti partecipativi di co-design, coinvolgendo *stakeholder* e possibili nuove audience nella fase di progettazione, per rispondere al meglio alle loro esigenze e fornire uno strumento realmente utile.

## PAROLE CHIAVE

Digital archive; human-centered design; design thinking; cultural heritage; co-design

## 1. INTRODUZIONE

Il progetto *PDS* (Piattaforma di Digital Storytelling) nasce con l'obiettivo di sviluppare una strategia per supportare la preservazione, la valorizzazione e la diffusione del patrimonio culturale materiale e immateriale conservato dall'Archivio Diaristico Nazionale (ADN), nel rispetto della peculiarità della scrittura manuale e della materialità del suo supporto. L'Archivio Diaristico Nazionale conserva infatti un fondo costituito da più di 9.900 testi autobiografici (si tratta principalmente di manoscritti e dattiloscritti - testimonianze, diari, epistolari - accompagnati in certi casi da immagini fotografiche, disegni, oggetti), consultabili solamente in loco su supporto cartaceo, e solo eccezionalmente a distanza, richiedendo una scansione parziale del testo via mail alla Direzione dell'Archivio. Nello specifico, la collezione dell'Archivio è composta da un fondo catalogato costituito da 9.412 diari (acquisiti dal 1985 al 2022) e una collezione di inediti da catalogare di 500 acquisizioni raccolte dal 2012 in poi. Nel complesso, a fine 2022, sono 9.912 i testi autobiografici raccolti da ADN, per un totale di 9.171 autori unici digitalizzati. Il progetto *PDS* sfrutta dunque le nuove tecnologie emergenti e le competenze culturali maturate attorno ai nuovi media per sviluppare soluzioni e strategie in grado di massimizzare l'accessibilità, l'usabilità e il riuso creativo dei materiali conservati. Il fine ultimo è quello di incrementare la partecipazione attiva dei portatori di interesse e della cittadinanza nei processi di raccolta e divulgazione dei testi autobiografici e di favorire un turismo culturale e sostenibile, intercettando nuovi pubblici e creando nuove audience. Il progetto *PDS* si pone

inoltre l'obiettivo di capitalizzare l'esperienza di progettazione fatta attraverso la creazione di un protocollo replicabile per la divulgazione digitale dei contenuti d'archivio, protocollo che include sia gli approcci progettuali *human-centered* (Jaillant, 2022) che gli aspetti giuridici legati alla tutela della privacy e dei diritti d'autore. Questo secondo obiettivo si concretizza in una seconda piattaforma digitale che raccoglie e condivide in open access quanto sopra descritto.

Il presente contributo si concentra sulla condivisione dell'esperienza *PDS*, attirando l'attenzione sulle metodologie di *human-centred design* e sugli strumenti partecipativi legati al processo di co-progettazione.

## 2. METODOLOGIA HUMAN-CENTERED

Il progetto di creazione della piattaforma di storytelling digitale per l'Archivio di Pieve Santo Stefano è stato concepito, fin dal principio, come un progetto basato sulla metodologia dello *human-centered design* (Brown, 2019) e sui mindset del *design thinking* (Rizzo, 2020). Lo *human-centered design* è un approccio alla progettazione e alla risoluzione di problemi basata sulle persone e, in particolar modo, sugli utenti a cui la progettazione si destina. È un tipo di progettazione che coinvolge la prospettiva umana e adotta il punto di vista degli utilizzatori ad ogni passaggio della progettazione e ad ogni fase di risoluzione dei problemi: comprensione dei bisogni e delle aspettative delle persone interessate, esplorazione delle possibili soluzioni, test di wireframe e prototipi. Non è un processo lineare, permette in ogni momento di ripercorrere le fasi precedenti per approfondire o chiarire gli aspetti rimasti irrisolti. Con l'applicazione di tale processo al progetto *PDS* si è cercato di rispondere alle domande: "In che modo le persone possono venire a conoscenza dell'esistenza dell'Archivio di Pieve Santo Stefano sul territorio nazionale e all'estero?"; "Come possono appropriarsi del valore della dimensione storica e narrativa del suo patrimonio anche a distanza?". Secondo questo approccio, le soluzioni efficaci e innovative si trovano all'intersezione di tre fattori: desiderabilità (utenti potenziali), sostenibilità (business) e fattibilità (implementazione tecnica). Dove questi tre fattori si sovrappongono, si ritiene che la soluzione sia auspicabile, finanziariamente sostenibile e tecnologicamente realizzabile (Brown, 2019). La ricerca di una soluzione rispetto alla domanda di partenza è stata sviluppata attraverso tre fasi principali: nella fase iniziale sono state ascoltate e comprese le esigenze dell'Archivio e le aspettative degli utenti attuali e potenziali. Gli strumenti utilizzati sono stati: analisi comparativa delle piattaforme di *storytelling* più interessanti ed efficaci in ambito culturale, interviste contestuali agli addetti ai lavori all'interno dell'Archivio Diaristico Nazionale, questionari e interviste agli utenti reali e potenziali identificati. Nella seconda fase sono state esplorate le possibili soluzioni rispetto all'architettura dell'informazione e sono stati realizzati i primi *sketch* della piattaforma in modo collaborativo (attraverso un processo di co-design), sketch che sono poi stati affinati grazie ai *feedback* ricevuti dalla rete di *stakeholder*. La terza e ultima fase è stata quella di testing ed è servita per validare il prototipo finale. Sono stati condotti dei test di usabilità e delle interviste per valutare l'esperienza utente. Le evidenze raccolte sono state consegnate sotto forma di report al team di sviluppo per procedere con l'implementazione finale della piattaforma. L'intera esperienza di progettazione (compresi metodologia e strumenti) è stata raccolta nel sito web dedicato e compone il *toolkit PDS*, un protocollo replicabile per la divulgazione digitale dei contenuti d'archivio.

## 3. L'IMPORTANZA DELLA FASE DI EMPATIZZAZIONE CON L'UTENTE

Analizzando più in dettaglio le varie fasi, dopo aver compreso il problema da risolvere e la sfida lanciata dal brief di design, il team dell'Università di Siena ha intrapreso la fase di user research per cercare di empatizzare con le persone a cui la piattaforma si destina. Dopo aver raccolto dati sugli utilizzatori attuali e aver individuato possibili nuovi utenti, sono state condotte numerose interviste. Inizialmente sono state ipotizzate cinque tipologie di possibili utenti e per ciascuna di esse le motivazioni (cosa li spinge/spingerebbe ad usare *PDS*), i bisogni esistenti ancora non soddisfatti, gli obiettivi e le aspettative (cosa si aspettano di trovare sulla nuova piattaforma). Queste tipologie di utenti sono: gli insegnanti e gli studenti di scuole secondarie di primo e secondo grado, i ricercatori, gli amministratori di piattaforme editoriali online (per riflettere sulle funzionalità da includere nel back-end della piattaforma per facilitare il lavoro di coloro che la amministreranno) e una categoria trasversale a tutte le altre che è stata indicata come « gli amici dell'Archivio » e comprende i diaristi, i trascrittori di diari, i membri del comitato di lettura, e le quelle persone che ruotano attorno al patrimonio di Pieve Santo Stefano e sono fedeli alla mission dell'Archivio. Dopo aver individuato gli utenti potenziali, è stato necessario verificare i punti di contatto tra le varie tipologie e definire quali fossero i bisogni comuni e le opportunità da cogliere per muovere verso una possibile proposta finale che rispondesse alle loro aspettative. Per ogni categoria individuata è stato selezionato un

campione di utenti reali (es. un campione di ricercatori, un campione di insegnanti, ecc.) e si è proceduto con l'organizzazione di interviste contestuali, interviste telefoniche e la somministrazione di questionari, attraverso i quali raccogliere dati e confermare le diverse categorie ipotizzate. Tale processo ha permesso di confermare gli utenti ipotizzati e di definire delle *personas*: degli archetipi di utenti con le aspettative, i bisogni, gli obiettivi specifici del gruppo che rappresentano. Per ciascuna *persona* ideata sono state indicate delle caratteristiche biografiche, l'occupazione, ma anche le abitudini di lettura o di ascolto su supporto digitale. Si sono subito delineate alcune tendenze rispetto alle abitudini e ai supporti di lettura, alle pratiche digitali rivolte alla didattica e alla necessità della consultazione di documenti a distanza nell'ambito della ricerca. Rispetto alle previsioni iniziali, alcune funzionalità sono risultate superflue mentre altre, non ipotizzate, si sono affermate come necessarie. In particolare, è emersa l'importanza della dimensione didattica e si sono organizzati focus group di co-progettazione con insegnanti di tutti i gradi per riflettere all'utilizzo dei testi autobiografici in classe e ai possibili scenari d'uso della piattaforma PDS all'interno delle scuole: sia per le scuole secondarie di primo grado che per le scuole secondarie di secondo grado è risultato interessante poter utilizzare le testimonianze dell'Archivio Diaristico Nazionale come fonti per l'insegnamento della storia e come supporto per l'insegnamento della letteratura (e in particolare dei moduli dedicati alla scrittura autobiografica). Attraverso le testimonianze autobiografiche è inoltre possibile creare percorsi didattici interdisciplinari indirizzati a materie come arte e educazione civica. In generale, insegnanti e studenti hanno sottolineato come i libri di testo non riportino documenti sulla vita dei cittadini comuni nelle diverse epoche storiche e i diari potrebbero essere davvero preziosi in questo senso, perché danno accesso a degli spaccati di vita e sono particolarmente efficaci nel permettere agli studenti di identificarsi, di sentire epoche storiche passate più vicine e di essere dunque maggiormente coinvolti nello studio della materia. È inoltre importante, e anche questo è stato rilevato sia dai docenti che dagli studenti, lasciarsi trasportare dalla potenza evocativa di questi testi (e accompagnarli con elementi iconografici, audiovisivi, sonori). Tutte le evidenze emerse dalle interviste e dai focus group sono state fondamentali per definire i contenuti e le funzionalità principali della piattaforma e fornire dei tentativi di risposta più adatti alle esigenze reali degli utenti cui la piattaforma si destina. In particolare, questa fase della ricerca ha permesso di definire meglio le abitudini di ogni utente e di coglierne in modo più preciso le esigenze e gli obiettivi nell'uso di PDS. Questo ha portato il team a decidere di declinare la piattaforma in tre diversi livelli d'accesso, per rispondere a tre diverse modalità di fruizione e a tre diversi tipi di esigenze, dunque, di fatto, a tre diverse tipologie di utenti: 1° livello, vetrina dell'archivio per tutti; 2° livello, strumento didattico; 3° livello, strumento di ricerca (digital library).

#### 4. IL PROCESSO DI CO-DESIGN

Dopo aver definito i problemi degli utenti che la nuova piattaforma dovrebbe risolvere, sono state ipotizzate le prime soluzioni, descritte come possibili scenari d'uso, sotto forma di *storyboards*. Per passare dalla fase di esplorazione delle prime soluzioni alla progettazione della soluzione finale sono state organizzate delle sessioni di co-design. In questa fase è stata dedicata particolare attenzione allo scenario del secondo livello di accesso alla piattaforma, interamente dedicato alla didattica. Per la scelta di architettura, funzionalità e contenuti sono stati realizzati dei *workshops* con diversi gruppi di lavoro nell'intento di raccogliere esigenze e desiderata rispetto ai contenuti e ai percorsi da proporre a insegnanti e studenti, con l'obiettivo di rendere la piattaforma un vero e proprio strumento per la didattica, efficiente e modulabile a seconda delle esigenze dei diversi gradi scolastici. È stata così costituita una rete di insegnanti di italiano, storia e geografia di scuole primarie e secondarie di primo e secondo grado, provenienti da tutta Italia, per riflettere insieme a loro sulle esigenze specifiche in materia di contenuti, percorsi editoriali ed ergonomia della navigazione per l'utilizzo della piattaforma a supporto dell'insegnamento. Durante i *workshops* si è cercato di raccogliere feedback sia sull'esperienza attuale degli insegnanti rispetto alla condivisione di contenuti digitali in classe e ai supporti di condivisione utilizzati, sia sui desiderata rispetto alla nuova piattaforma PDS. Per rispondere alla prima domanda si è cercato di comprendere in che modo gli insegnanti condividono in classe i contenuti digitali con gli studenti; di quali supporti si avvalgono per presentare loro gli esercizi e se svolgono attività di lettura collettiva o individuale in classe a partire da supporti digitali. Le testimonianze dei soggetti sono state argomentate e discusse con gli altri partecipanti. Per raccogliere desiderata e aspettative riguardo la nuova piattaforma sono invece state organizzate alcune attività di co-design tra cui: ordinare i contenuti tematici in ordine di importanza a fini didattici (dal maggiore al minore interesse per la classe); indicare percorsi tematici (grandi guerre, amicizia, oggetti, altro) e motivarne la scelta; definire la linea temporale desiderata per la navigazione dei contenuti (anno, secolo, grandi eventi, altro); creare un esercizio

didattico sulla base dei contenuti dell'archivio; svolgere un esercizio tra quelli proposti dagli altri partecipanti al workshop; discutere degli esercizi proposti dai colleghi e della loro validità, fattibilità, difficoltà. Le attività elencate, insieme ad altri momenti di confronto, hanno messo in luce alcune esigenze specifiche dei soggetti coinvolti, come ad esempio la possibilità di poter utilizzare le testimonianze conservate dall'Archivio appoggiandosi a percorsi "pronti all'uso" che si focalizzano su contenuti e tematiche precise, in modo da evitare dispersione e perdite di tempo durante la lezione. Gli insegnanti suggeriscono, tra i percorsi interessanti per gli studenti, un percorso sui diari della Prima Guerra Mondiale (la guerra di trincea attraverso gli stati d'animo delle persone che l'hanno vissuta). Questo permetterebbe di scoprire battaglie intime, personali, che permetterebbero di avvicinare il vissuto dei protagonisti della storia a quello degli studenti). I testi dell'Archivio sulla Grande Guerra potrebbero inoltre essere affiancati dall'insegnamento del medesimo periodo in materie come storia e letteratura (es. Ungaretti). Un'altra esigenza emersa è stata quella di accompagnare i testi con immagini e contenuti multimediali (audio, video) per favorire l'apprendimento e l'accessibilità dei diari (Roncaglia, 2018). In seguito a tali richieste è apparsa con maggiore chiarezza la necessità di stringere relazioni con altri archivi (fotografici e cinematografici). È stato avviato un primo partenariato con *Home Movies*, l'Archivio dei Film di Famiglia di Bologna. Insieme a *Home Movies* è stato concepito un percorso tematico che potesse essere comune ai due archivi e metterli in dialogo: *Lungo la via Emilia*. Si tratta di un percorso che contiene diari testuali emiliani e romagnoli conservati dall'Archivio Diaristico Nazionale, illustrati da immagini filmiche di diari visivi emiliani e romagnoli conservati da *Home Movies*. Quest'ultimo esempio evidenzia come il processo di co-design abbia guidato passo a passo i designer nel passaggio dal wireframe (la rappresentazione geometrica della struttura e dell'architettura dell'informazione) alla creazione di un prototipo - a media fedeltà su Adobe XD - che prevedesse anche le interazioni e una prima visualizzazione grafica della piattaforma vicina al risultato finale, così da tenere in conto anche la dimensione iconografica e la riflessione rispetto all'uso della piattaforma finalizzata.

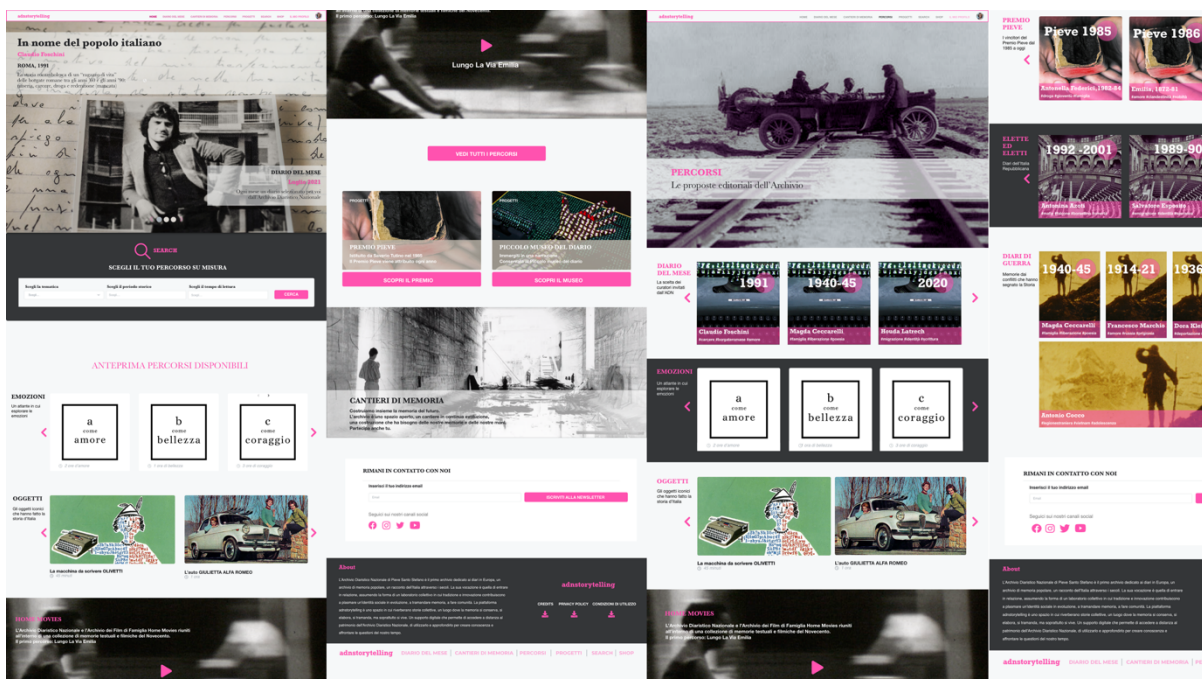


Figura 1. Da sinistra verso destra: home alta, home bassa; pagina percorsi alta, pagina percorsi bassa.

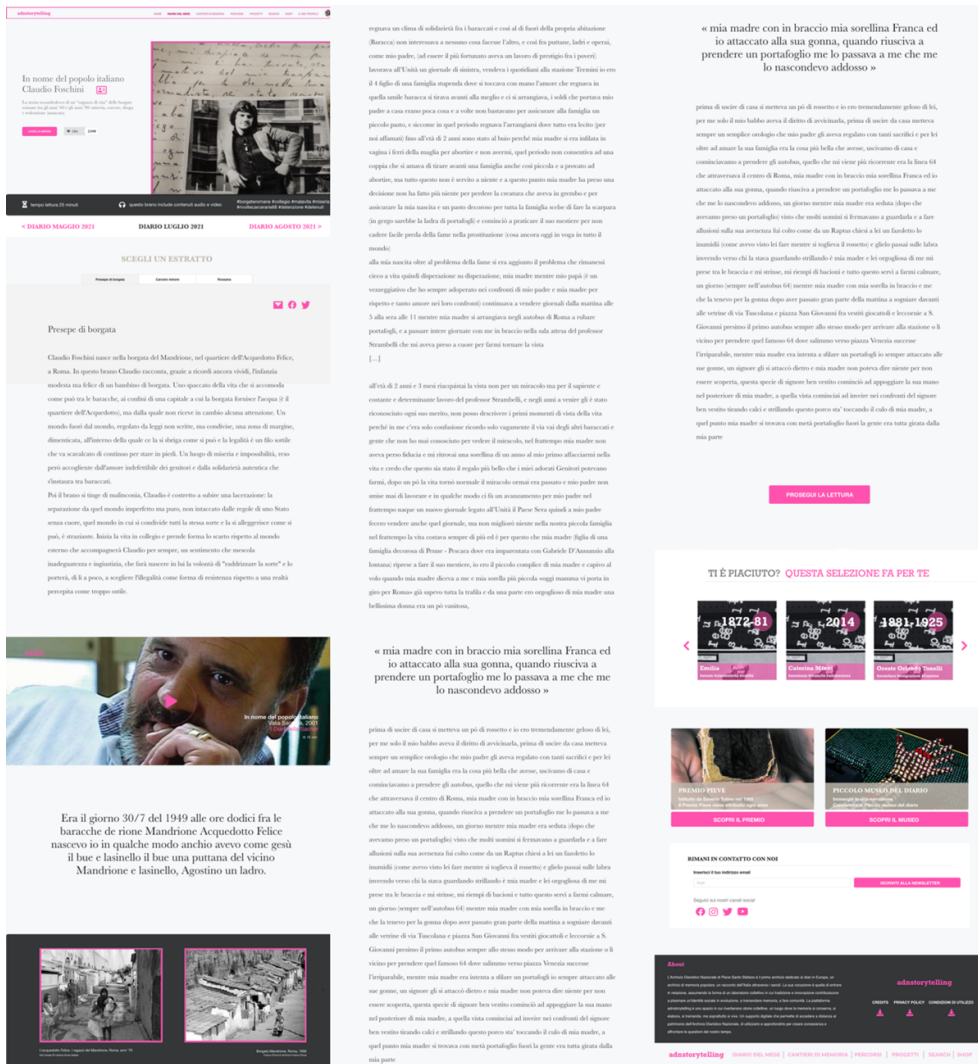


Figura 2. Collage orizzontale della visione in lunghezza di un diario del prototipo interattivo realizzato con Adobe XD.

## 5. RISULTATI

I risultati di questo processo di co-design si sono concretizzati in due output: un prototipo interattivo della piattaforma PDS co-progettato e validato dagli utenti e un sito web contenente il toolkit PDS, le linee guida per replicare la metodologia di progettazione in nuovi contesti di storytelling di digital memories. L'approccio human-centered ha permesso di progettare la piattaforma insieme alle persone a cui si destina e di validare passo a passo la struttura, le funzionalità e i contenuti della piattaforma, nonché di creare una rete di futuri utenti fin dalla fase di progettazione. Il prototipo finale, realizzato con il software XD di Adobe, rappresenta un modello interattivo della piattaforma di storytelling, cucito sulla base delle esigenze e delle aspettative emerse dalle interviste, dai focus-group, dai workshop di co-progettazione e dai continui riscontri raccolti da utenti e stakeholder. Questo modello è stato testato da soggetti selezionati perché rappresentativi delle varie audience individuate (pubblico generico, insegnanti/studenti, docenti e ricercatori) e il report dei risultati è stato condiviso con il gruppo di sviluppo per procedere all'implementazione del servizio finale. Il report contiene i risultati dei test A/B (preferenza tra due soluzioni di interfaccia diverse per la stessa tipologia di contenuto) e le osservazioni dell'esperienza utente valutate tramite metodi quantitativi (tempi di esecuzione dei compiti e numero di obiettivi raggiunti) e qualitativi (come il processo think aloud in cui il soggetto commenta ad alta voce l'esperienza in corso), allo scopo di comprendere criticità e opportunità delle soluzioni proposte. Al termine di ogni test ai partecipanti è stato fatto compilare un questionario di valutazione dell'esperienza, realizzato internamente, e un questionario di usabilità basato sul System Usability Scale (Lewis, 2018). Tutti i risultati delle analisi svolte e tutti i passaggi operativi che sono stati affrontati per arrivare alla strutturazione del prototipo finale sono raccolti nel toolkit PDS, che prende la



forma di un sito web: <https://sites.google.com/unisi.it/toolkitpds/home>. Il toolkit del progetto PDS è stato suddiviso in rubriche.

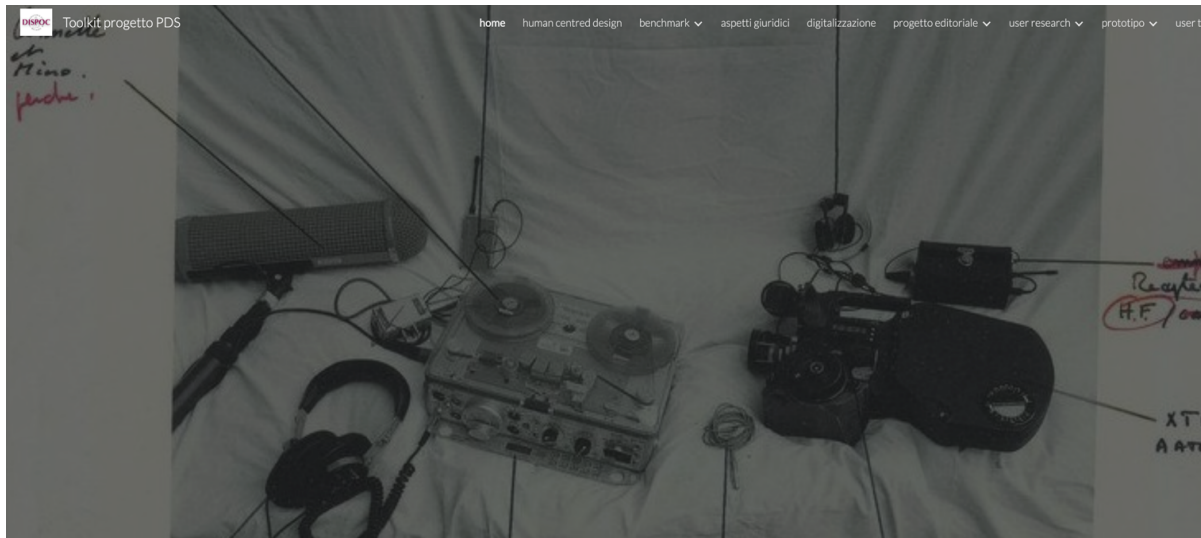


Figura 3. Pagina home del toolkit *PDS*, visione della header e del menù principale.

Ogni rubrica corrisponde a una fase di analisi o a un passaggio operativo. È possibile orientarsi attraverso il menù principale del sito e accedere alle informazioni leggendo i contenuti online e scaricando i documenti utili (tra cui anche le possibili licenze e i documenti giuridici redatti per regolamentare la condivisione online dei documenti archiviati). La volontà dell'Archivio Diaristico Nazionale e della squadra di design dell'Università di Siena è quella di condividere sia i risultati che le operazioni necessarie alla costruzione di una piattaforma di storytelling di valorizzazione di un patrimonio archivistico, con l'augurio che questo strumento online, aperto e accessibile a tutti, possa essere utilizzato a vantaggio di altri progetti di valorizzazione del patrimonio storico e culturale del nostro paese. La prospettiva futura è quella di affinare il toolkit e usarlo come guida per costituire un vero e proprio ecosistema digitale per garantire l'accesso al patrimonio culturale a tutte le possibili categorie di fruitori, integrando la piattaforma *PDS* con vari sistemi aperti. Tra questi sarà prevista l'implementazione di una *digital library*, un contenitore digitale per la conservazione di tutte le unità archivistiche che costituiscono il fondo ADN sotto forma di PDF percorribili, con tutti i livelli informativi associati (catalogazione, metadattazione, allegati, trasposizioni audio).

## 6. RINGRAZIAMENTI

Si ringraziano il Professor Tiziano Bonini Baldini, tutor scientifico e ideatore del progetto; Natalia Cangì, Direttrice dell'Archivio Diaristico Nazionale di Pieve Santo Stefano e i suoi collaboratori Filippo Massi e Massimiliano Bruni per la generosa guida alla scoperta del patrimonio dell'Archivio Diaristico Nazionale e la preziosa collaborazione durante tutte le fasi della ricerca; il Dipartimento di Scienze Sociali Politiche e Cognitive dell'Università di Siena; Roberto Ferrari, Direttore della Direzione *Cultura e Ricerca* Regione Toscana (2015-2020), che ha reso possibile dedicare due anni a questa ricerca attraverso l'iniziativa *100 ricercatori per la cultura*, all'interno della quale è stato bandito l'assegno di ricerca attribuito a Martina Magri.

## BIBLIOGRAFIA

- [1] Brown, Tim. *Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation*. New York: HarperCollinsPublishers, 2019.
- [2] Marras, Cristina, Marco Passarotti, Greta Franzini, and Eleonora Litta. "Atti Del IX Convegno Annuale AIUCD. La Svoltta Inevitabile: Sfide e Prospettive per l'Informatica Umanistica." Text, 2020. <https://doi.org/10.6092/UNIBO/AMSACTA/6316>
- [3] Jaillant, Lise. "Chapter 3: Design Thinking, UX and Born-Digital Archives: Solving the Problem of Dark Archives Closed to Users." In *Digital Humanities Research*, edited by Lise Jaillant, 1st ed., 2:83–108. Bielefeld, Germany: Bielefeld University Press / transcript Verlag, 2022. <https://doi.org/10.14361/9783839455845-004>
- [4] Lewis, James R. "The System Usability Scale: Past, Present, and Future." *International Journal of Human-Computer Interaction* 34, no. 7 (July 3, 2018): 577–90. <https://doi.org/10.1080/10447318.2018.1455307>

- [5] Tomasi, Francesca. "DIGITAL HUMANITIES AND DIGITAL KNOWLEDGE (DHDK) International Second Cycle/Master Degree." *Umanistica Digitale*, May 22, 2018, No 2 (2018). <https://doi.org/10.6092/ISSN.2532-8816/7862>
- [6] Tomasi, Francesca. *Organizzare La Conoscenza: Digital Humanities e Web Semantico: Un Percorso Tra Archivi, Biblioteche e Musei*. *Biblioteconomia e Scienza Dell'informazione* 39. Milano: Editrice Bibliografica, 2022. Z666.5.T66 2022.
- [7] Tomasi, Francesca. "AIUCD. Un portale concettuale per le Digital Humanities in Italia." *Griseldaonline*, December 15, 2021, 205-212 Paginazione. <https://doi.org/10.6092/ISSN.1721-4777/12375>
- [8] Rizzo, Antonio. *Ergonomia cognitiva. Dalle origini al design thinking*. Bologna: Il Mulino, 2020.
- [9] Roncaglia, Gino. *L'età della frammentazione: cultura del libro e scuola digitale*. I Robinson. Bari: Laterza, 2018. 303.483 3.
- [10] Weston, Paul Gabriele, Baldini, Primo, Carbé, Emmanuela, Pusterla, Laura. "Archivi digitali di persona. PAD – Pavia archivi digitali e gli archivi degli scrittori". *DigItalia*, XIV, 1, 2019, 31-54. <https://digitalia.cultura.gov.it/article/view/2274/1545>



# Un futuro digitale

## per gli archivi de ‘Gli Anni della Cupola’

Manuel Favaro<sup>1</sup>, Eva Sassolini<sup>2</sup>, Alessandra Cinini<sup>3</sup>, Margaret Haines<sup>4</sup>

<sup>1</sup>Istituto di Linguistica Computazionale ‘A. Zampolli’ (CNR-ILC), Italy - manuel.favaro@ilc.cnr.it

<sup>2</sup>Istituto di Linguistica Computazionale ‘A. Zampolli’ (CNR-ILC), Italy - eva.sassolini@ilc.cnr.it

<sup>3</sup>Istituto di Linguistica Computazionale ‘A. Zampolli’ (CNR-ILC), Italy - alessandra.cinini@ilc.cnr.it

<sup>4</sup>Opera di Santa Maria del Fiore di Firenze, Italy- mhaines0@gmail.com

### ABSTRACT

L’articolo si inserisce tra le iniziative di recupero, salvaguardia e valorizzazione di archivi digitali conservati in formati non standard che l’Istituto di Linguistica Computazionale “A. Zampolli” sta promuovendo da alcuni anni con diverse strategie legate alla natura delle risorse testuali trattate [8, 9 e 10]. In queste iniziative l’istituto offre inoltre la possibilità di depositare gli output sull’infrastruttura internazionale CLARIN<sup>1</sup>, al fine di garantirne la preservazione a lungo termine e la divulgazione. In questo particolare caso sono discussi i metodi, le criticità e le strategie di conversione di dati digitali strutturati e codificati con un sistema software proprietario, in un nuovo formato di rappresentazione in linea con gli standard internazionali TEI. Il contenuto è costituito dall’edizione completa delle fonti amministrative conservate presso l’Opera di Santa Maria del Fiore di Firenze relative al periodo dal 1417 al 1436 quando fu progettata e costruita la Cupola di Brunelleschi. L’intervento intende proporre una riflessione metodologica sulla standardizzazione di edizioni digitali obsolete, la conservazione a lungo termine e sulle problematiche legate alla condivisione dei dati con la comunità scientifica. Gli archivi hanno infatti un interesse multidisciplinare, non solo per le informazioni storiche e artistiche che vi sono contenute, ma anche per le peculiarità che riguardano la lingua in cui gli atti originali sono stati redatti: una delle più antiche e autentiche attestazioni del volgare fiorentino, usato non con intenti letterari, ma pratici. Il processo di conversione riguarda tutti i materiali digitali, sia i metadati inseriti manualmente durante la preparazione dei testi, sia il contenuto originale degli atti, e mira a rappresentare opportunamente l’edizione digitale dei testi, ma vuole anche conservare l’apparato critico predisposto dai redattori che correda le schede di voci analitiche sulle quali si basano le funzioni di ricerca all’epoca realizzate.

### PAROLE CHIAVE

Digital Humanities; TEI mapping; corpora; knowledge discovery/representation

### 1. INTRODUZIONE

Nel vasto panorama degli archivi digitali, le collezioni di fonti primarie sono ormai numerose. Anche l’Italia possiede molteplici raccolte di risorse digitali provenienti da manoscritti e da incunaboli, sempre più multimediali, che offrono all’utente ampie possibilità di accesso e di studio dei materiali presenti in rete.

L’archivio digitale “Gli Anni della Cupola” rientra appieno in questo universo, in quanto raccoglie la documentazione dell’Opera di Santa Maria del Fiore di Firenze per la Cupola del Brunelleschi nel suo ricco contesto istituzionale. I documenti, scritti sia in latino, sia in volgare, in qualche caso anche misti, testimoniano i modi con cui l’Opera gestiva il rifornimento e il trasporto dei materiali edilizi, qual era l’organizzazione del personale interno ed esterno, come venivano amministrate le finanze, e molte altre informazioni, non soltanto legate al cantiere della cupola, ma anche riguardanti gli arredi e le opere d’arte, gli immobili del clero, persino le altre mansioni esterne affidate all’Opera dal Comune di Firenze [1 e 5]. L’archivio comprende oltre 21000 atti trascritti con criteri filologici in formato digitale proprietario a partire dal 1996 grazie alla collaborazione tra l’Opera e l’Istituto di Linguistica computazionale “A. Zampolli” del CNR di Pisa (ILC). Il progetto di diffusione degli archivi digitali vede anche la collaborazione del Max-Planck Institut per la Storia della Scienza di Berlino che a partire dal 2001 ne ha prodotto un’edizione codificata in linguaggio HTML, consultabile online<sup>2</sup>, precoce esempio di edizione digitale open access per la storia della tecnologia.

L’ingente mole di dati rende l’archivio prezioso e del tutto particolare, ricco di informazioni testuali, metatestuali, paratestuali ed editoriali; tuttavia, col passare degli anni, sta inevitabilmente andando incontro a una rapida obsolescenza.

<sup>1</sup> Infrastruttura europea per creare, coordinare e rendere le risorse linguistiche e le tecnologie disponibili e prontamente utilizzabili: [www.clarin.eu](http://www.clarin.eu)

<sup>2</sup> <http://duomo.mpiwg-berlin.mpg.de/>

Si è reso perciò necessario adottare opportune strategie per recuperare i dati da quel formato digitale non standard e portarlo verso un formato di rappresentazione XML-TEI, che oggi costituisce per molte iniziative la migliore garanzia di sopravvivenza<sup>3</sup>. Era obiettivo del progetto anche attuare strategie di conservazione a lungo termine oggi molto importanti, che vedono di grande attualità il ruolo delle infrastrutture di ricerca per le Scienze Umane e Sociali (SSH).

## 2. STUDIO DEL MODELLO DI RAPPRESENTAZIONE

Il recupero e la salvaguardia dell'archivio parte, in primo luogo, dallo studio dei dati e dal relativo formato. Il progetto di digitalizzazione è iniziato nella seconda metà degli anni '90 del secolo scorso, ed è quindi legato alle tecnologie del tempo, quando non era ancora consolidato un formato standard di rappresentazione. Le prime attività furono avviate dall'Opera sotto la direzione di Margaret Haines, storica dell'arte specializzata nella documentazione sul Duomo di Firenze, che ha seguito tutte le fasi del progetto. Il sistema adottato per l'immissione dati (la stazione lessicografica "WSLexx"<sup>4</sup>) è un software proprietario che era collegato a un sistema di indicizzazione dei dati e al relativo sistema di analisi testuale (DBT: *Data Base Testuale*<sup>5</sup>) per la consultazione della banca dati. Nella fase di costruzione dell'archivio digitale le competenze che ILC ha messo a disposizione sono state operativamente indirizzate alla codifica delle schede nella banca dati, ma anche alla revisione dei dati inseriti attraverso una indicizzazione molto complessa. Il successivo apporto del Max-Planck Institut è stato fondamentale invece per la visibilità e fruibilità dei dati al grande pubblico, perché ne ha prodotto una versione Web, a partire da una pionieristica conversione dei dati in SGML-XML con relativo schema ODD. Oggi anche quest'ultima risorsa mostra i suoi limiti poiché *document oriented*, ovvero segue uno schema proprio, non derivato da alcuno standard di rappresentazione; situazione che rende l'archivio lontano da criteri condivisi di rappresentazione e non interoperabile con altre risorse simili. Il ricorso all'impiego dello strumento di marcatura più adoperato da chi pubblica materiale testuale in rete, ovvero l'HTML, è destinato alla lunga a non poter soddisfare esigenze di conservazione e prospettive certe di riconversione del materiale senza perdita rilevante di informazioni. Inoltre, quella rappresentazione risente del legame con le funzionalità del sito Web per le quali era stata pensata rendendo difficile abbinare all'opzione digitale un reale, concreto miglioramento dei risultati delle ricerche che vi si possono implementare.

La risorsa da cui siamo partiti era già un'edizione digitale e il mancato allineamento con gli attuali standard di catalogazione archivistica era una caratteristica prevedibile data la natura particolare dei dati. In aggiunta, l'intervento dell'ILC non era finalizzato a un eventuale processo di conversione verso tali standard, ma principalmente al recupero e al salvataggio dei testi. La letteratura di settore indica strade diverse a seconda del tipo d'intervento [3], che nel nostro caso ha suggerito l'adozione di uno standard di rappresentazione di testi. Da qui la scelta di costituire un modello dei dati XML-TEI [11], seguendo in primo luogo le linee guida P5<sup>6</sup>, studiando la letteratura di settore più vicina alle nostre esigenze [2, 4 e 6] e avvalendosi del confronto con risorse che hanno cercato di superare i limiti di TEI nella rappresentazione di fonti primarie, come il sottoinsieme di tag fornito da EpiDoc<sup>7</sup>, volto alla rappresentazione di documenti digitali antichi, anche provenienti da manoscritti. TEI, infatti, è considerato lo standard *de facto* per la codifica dei testi [4], in quanto si propone di contemplare tutti i fenomeni di interesse umanistico; tuttavia, avendo l'obiettivo di fornire un vocabolario unico di elementi non ambigui, restano ancora problemi sull'esaustività dei fenomeni ricompresi, specialmente in casi peculiari come quello trattato, dove l'espressività degli interventi era ottenuta anche ricorrendo al software di indicizzazione. Ne sono la prova le numerose insidie che si distribuiscono a vari livelli e accompagnano il lavoro di conversione. In primo luogo, le informazioni presenti nei metadati sono state inserite manualmente e ricavate direttamente dai testi nella forma in cui occorre, da cui deriva un elevato tasso di variabilità dei contenuti: per esempio, nel campo 'nomi e qualifiche', la qualifica *abate* può ritrovarsi da sola, accompagnata dalla specificazione, più o meno dettagliata, del luogo di competenza (*abate della badia di Firenze*, o *abate di Firenze*), oppure da precisazioni che puntualizzano il ruolo del soggetto in questione rispetto al documento (*abate debitore*). In secondo luogo, il gran numero di campi in cui sono stati organizzati i metadati non sempre trova un preciso corrispettivo nello standard TEI, dove i nomi dei tag hanno un valore semantico diverso da quello inteso dagli estensori delle schedature: il campo 'iconografia', ad esempio, non riguarda la descrizione delle decorazioni (che in TEI corrisponde al tag <decoDesc>), ma fa esplicito riferimento a dati iconografici menzionati all'interno dello specifico atto; un ulteriore esempio è la stessa qualifica, accennata poc'anzi, che nella logica dei redattori/schedatori riguardava anche mansioni e mestieri (p.e. *fabbro*).

---

<sup>3</sup> <https://digiliblt.uniupo.it/index.php>

<sup>4</sup> Strumento software realizzato da Eugenio Picchi che nasce dalle ricerche svolte presso CNR-ILC sulle prime procedure informatiche per la costruzione di dizionari.

<sup>5</sup> Motore di analisi testuale anch'esso ideato da Eugenio Picchi.

<sup>6</sup> <https://tei-c.org/Guidelines/P5>

<sup>7</sup> <https://epidoc.stoa.org/>

Tutte queste informazioni, frutto del lavoro di anni da parte dello staff del progetto di trascrizione, sono irrinunciabili, ma rendono il lavoro di conversione più complesso, proprio perché il punto di partenza non è solo una raccolta testuale, ma un'edizione digitale composta da singoli atti di diverse tipologie, organizzata secondo criteri che non avevano intenzione di regolarizzare i dati, ma di renderli in stretta aderenza ai testi originali, normalizzando soltanto l'ortografia in stile moderno. Ciò non permette di seguire scelte di codifica già utilizzate in letteratura da altri, come per esempio il progetto di edizione digitale del Codice Pelavicino, importante manoscritto del XII secolo, curato dal Laboratorio di Cultura Digitale dell'Università di Pisa<sup>8</sup>; tale progetto si basa infatti su una quantità selezionata di informazioni da codificare, riuscendo nell'intento di normalizzare i dati e di consentire l'interrogazione dei testi anche sulla base delle entità considerate più rilevanti estratte automaticamente [7].

Lo studio e la ricerca di strategie che permettano la mediazione tra i contenuti dell'archivio, distribuiti a diversi livelli, e le linee guida dello standard TEI, è un processo che non si è ancora concluso. In una prima fase si è definita una struttura generale, che sfruttasse le funzionalità del <teiCorpus> e includesse: un <teiHeader> contenente i metadati generali dell'edizione digitale; le schede, sotto l'elemento <TEI>, comprendenti i <teiHeader> in cui codificare i metadati estratti dai campi, e i testi delle trascrizioni (<text>), che a loro volta accolgono anche l'eventuale titolo originale dell'atto e l'insieme del paratesto, ossia le note e i riferimenti bibliografici.

Particolare attenzione è stata prestata alla ricerca del miglior modello per la conversione dei campi. L'eterogeneità delle informazioni da trattare e la loro diversa strutturazione ha condotto alla scelta di raggruppare le informazioni in liste, nella gran parte dei casi generiche (<list>); dove però si poteva optare per tag più specifici, si è ricorso a <listPerson> per il campo riguardante la codifica dei nomi e delle qualifiche, <listPlace> per i campi riguardanti i luoghi citati nel testo, <listObject> per gli oggetti e per i materiali.

Questo primo tentativo di codifica ha permesso di ipotizzare le regole per la codifica automatica di tutte le schede dell'archivio e di eseguire delle prove del modello verificando casi problematici ed eccezioni impreviste. Tuttavia, una valutazione dei risultati ottenuti convertendo tutto l'archivio, secondo tale modello, ha mostrato problemi di completezza e comprensione dei dati. La scelta di inserire nell'header delle singole schede tutte le informazioni a corredo di ogni atto andava a discapito della leggibilità degli output, apparendo ridondanti. Per questo motivo tale rappresentazione, pur ricalcando la classica struttura del <teiCorpus>, non permetteva la descrizione dei 30 manoscritti da cui provengono gli atti, poiché era difficilmente collocabile in questo spazio; una condizione che ha imposto una revisione del modello, una modifica operata con l'intenzione di evitare di raccogliere tali informazioni esternamente, ma anche di non sovraccaricare gli header delle singole schede con informazioni ripetute. Si è quindi optato per un modello più articolato che prevede un ulteriore livello <TEI> intermedio dedicato alla rappresentazione delle unità archivistiche. In definitiva, quindi, il <teiCorpus> del progetto contiene: il <teiHeader> con i metadati generali; un livello <TEI> che contiene un ulteriore <teiHeader> con la descrizione dei manoscritti come unità archivistiche; un altro livello <TEI> dove sono codificati i campi e i testi riferiti agli atti (lo schema del modello è riportato in Figura 1).

---

<sup>8</sup> <http://www.labcd.unipi.it>

```

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE teiCorpus SYSTEM "tei_ms.dtd">

<teiCorpus><!--contiene tutto l'archivio-->
  <teiHeader><!--intestazione del corpus (metadati relativi all'intero archivio)--></teiHeader>

  <TEI><!-- inizio del manoscritto n.1-->
    <teiHeader><!--contiene l'intestazione del manoscritto n.1--></teiHeader>

    <TEI><!--inizio della scheda n.1(1)-->
      <teiHeader><!--contiene l'intestazione/campi della scheda n.1(1)--></teiHeader>
      <text><!--contiene il testo originale della scheda n.1(1)--></text>
    </TEI><!--fine della scheda n.1(1)-->

    <TEI><!--inizio della scheda n.2(1)-->
      <teiHeader><!--contiene l'intestazione/campi della scheda n.2(1)--></teiHeader>
      <text><!--contiene testo originale della scheda n.2(2)--></text>
    </TEI>
    ...
    <!-- seguono tutte le schede del manoscritto n. 1-->
    ...
  </TEI><!--fine manoscritto n. 1-->

  <TEI><!-- inizio del manoscritto n.2-->
    <teiHeader><!--contiene l'intestazione del manoscritto n.2--></teiHeader>

    <TEI><!--inizio della scheda n.1(2)-->
      <teiHeader><!--contiene l'intestazione/campi della scheda n.1(2)--></teiHeader>
      <text><!--contiene testo originale della scheda n.1(2)--></text>
    </TEI><!--fine della scheda n.1(2)-->
    ...
    <!-- seguono tutte le schede del manoscritto n. 2-->
    ...
  </TEI><!--fine manoscritto n.2-->
  ...
  <!-- seguono tutti i manoscritti, ognuno con le schede di competenza-->
  ...
</teiCorpus>

```

Figura 1. Schema del modello di rappresentazione XML-TEI.

### 3. STRUTTURAZIONE DEI CONTENUTI E STRATEGIE DI CONVERSIONE AUTOMATICA

Una ulteriore complessità che accompagna questo intervento di recupero riguarda l'ingente mole dei dati. Infatti, mentre per l'organizzazione delle informazioni da inserire nell'header generale del corpus, così come per le descrizioni dei 30 manoscritti, si è deciso di optare per una compilazione manuale, per le oltre 21000 schede e il relativo corredo di metadati è stato necessario individuare un approccio automatico alla codifica. Non potendo contare su una banca dati strutturata come un database, le normali procedure di estrazione attuabili in quei casi erano precluse; inoltre, anche gli approcci alla strutturazione basati su algoritmi di NLP, come le tecniche di Named Entity Recognition, non trovano adeguata applicazione, date le peculiarità della lingua (sia volgare, sia latino) e dei criteri di resa di nomi, luoghi, materiali, ecc. nei rispettivi campi. L'approccio infine scelto si rifà a precedenti esperienze condotte da ILC sul recupero di risorse da formati non standard [8] e mira a produrre una codifica automatica attraverso un 'processo di parsing' (ovvero di analisi sequenziale automatica) dell'input. In generale, in informatica tale processo indica il riconoscimento della struttura implicita e/o esplicita di dati, che vengono segmentati in frammenti, sulla base di una 'grammatica formale' che ne identifica gli elementi qualificanti. Un parser è quindi un programma software che esegue tale compito usando la grammatica data.

Nel caso specifico di questi archivi, il lavoro di costruzione delle regole di riconoscimento e codifica dei dati deve assolvere al duplice compito di conservare l'organizzazione dei metadati e di codificare opportunamente il testo originale, sul quale sono presenti note, integrazioni redazionali segnalate da corsivi o da parentesi tonde e quadre. L'analisi puntuale dei dati ha mostrato come le difficoltà maggiori siano in larga parte legate al rispetto dello standard TEI: come già affermato nel paragrafo precedente, l'edizione prodotta rispondeva a specifiche esigenze, spesso legate agli strumenti utilizzati. Il

tentativo di normalizzazione, per quanto necessario, deve tenere conto della lunga gestazione del progetto, degli interventi manuali, delle scelte di edizione e del loro aspetto multidisciplinare. Sono presenti infatti campi i cui contenuti rivestono particolare interesse artistico, storico e linguistico: ne sono un esempio i contenuti dei campi oggetti, nomi e trasporti, sui quali sono stati condotti specifici studi [5].

Uno dei punti critici del processo di parsing è la sua notevole complessità: essendo condotto sequenzialmente su tutte le schede, le difficoltà riguardano sia gli algoritmi che interpretano le regole impostate, sia i tempi di esecuzione. Per agevolare l'elaborazione sono state studiate estrazioni preliminari di porzioni dei dati, in particolare di quei campi in cui serviva una maggiore articolazione in attributi e valori dei tag XML. Per esempio, la rappresentazione del campo <date> richiede la compilazione di attributi condizionali a seconda che si sia in presenza di date certe, multiple, intervalli aperti o chiusi. Gli archivi presentano tutte queste varianti ed è stato quindi necessario pre-elaborare i dati delle schede per quel campo e costruire per esso una risorsa esterna, già pronta all'uso in fase di parsing. Una strategia simile è stata utilizzata anche per il campo <locus> (carta), per il quale avevamo più tipologie di rappresentazione: carta singola, doppia (recto/verso), fascicolo o intervallo. Questo tipo di intervento di estrazione ed elaborazione dei contenuti si è tuttavia limitato prevalentemente ai metadati, mentre per la parte testuale e originale dei documenti si è operato nel pieno rispetto dei dati, senza modificare gli interventi fatti in fase di trascrizione.

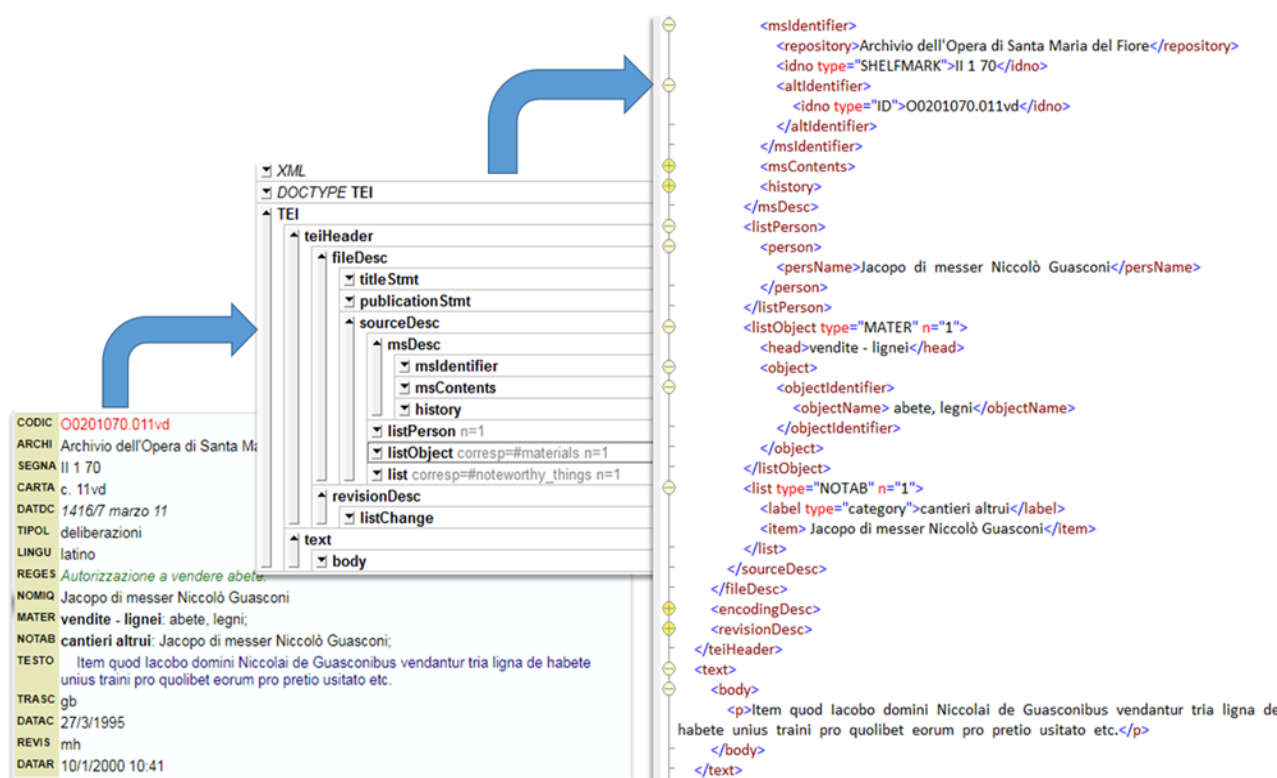


Figura 2. Esempio di conversione di scheda d'archivio, da versione HTML a XML- TEI.

#### 4. PROSPETTIVE DI UTILIZZO/SFRUTTAMENTO DEI DATI CONVERTITI

Nel preservare il patrimonio culturale è buona norma occuparsi non solo della conservazione in un formato standard, ma anche della condivisione, ovvero di renderlo disponibile all'utente finale in modo che possa consultarlo in modo efficiente. Nel caso specifico, contemporaneamente alla conversione delle schede in formato XML-TEI è stata generata una prima struttura di database (in particolare un DB di tipo MySQL), che contiene le informazioni della quasi totalità dei campi predisposti dai redattori a corredo della parte testo di ciascuna scheda. In questa circostanza non si è trattato di pensare ad un modello di rappresentazione in DB dei dati, viste le criticità descritte, ma di individuare strategie per mappare automaticamente i contenuti dei metadati in tabelle e campi memorizzabili in forma di database, ammettendo una certa quantità di 'rumore'. Pur tenendo conto delle esigenze di razionalizzazione e di normalizzazione dei dati, che impone la costruzione di un DB, si è cercato dove possibile di rispettare le scelte adottate dalla redazione dell'archivio. Per questo motivo, trovano posto informazioni ridondanti in campi diversi e/o strutturazioni non sempre uniformi in campi quali 'oggetti', 'nomi' e 'trasporti'. Ad esempio, nel caso dei 'trasporti' per la categoria *dorso* troveremo specificato come mezzo sia la coppia di termini *soma, asino*, sia la sola parola *asino*; la scelta viene effettuata sulla base del testo della scheda e può quindi essere utilizzato un livello di dettaglio differente. Come per la codifica delle schede, anche per la costruzione



del DB sono state necessarie elaborazioni aggiuntive condotte sui dati dei campi meno descrittivi, come quello della data, la cui codifica, non essendo ‘parlante’, non poteva essere efficacemente utilizzata come utile filtro per le ricerche in fase di consultazione. Anche per le stringhe inserite nel campo ‘tipologia’ e quelle relative all’indicazione del numero di carta dell’atto di riferimento per la singola scheda sono state riscontrate alcune difformità inevitabilmente legate all’intervento umano nella trascrizione. I dati così ri-elaborati sono stati utilizzati sia a supporto della fase di conversione in formato XML per la compilazione dei valori e attributi degli specifici tag, sia per lo studio dei materiali e del modo di sfruttarli/usarli. Per esempio, le sintesi e le estrazioni di dati per il campo ‘materiali’ rappresentano una base di partenza per mappare i dati in ontologie note e condivise; si riscontrano infatti parole tipiche del territorio toscano, come nel caso dei materiali commestibili, o elenchi di termini più tecnici come le denominazioni dei diversi tipi di legname (anche poco noti).

Il complesso processo di conversione impostato pone inoltre le basi per una riflessione sul potenziamento futuro dello strumento di consultazione/studio a fini linguistici dei testi, che ha coinvolto il Dipartimento di Lettere e Filosofia dell’Università di Firenze (DILEF) nel 2016. Il progetto, nato come estensione de *Gli Anni della Cupola*, ha visto convergere gli interessi di carattere linguistico di ILC e DILEF, e ha prodotto una prima estrazione della banca dati dei soli campi testuali, a oggi consultabile online<sup>9</sup>, che potrebbe avvalersi delle potenzialità del DB realizzato in una prospettiva di consultazione più articolata, tale da integrare la ricerca testuale oggi disponibile. In questo scenario, è possibile sfruttare l’uso del DB come ulteriore accesso ai dati, in particolare producendo sintesi di vario tipo grazie all’organizzazione delle informazioni. In questo modo è possibile offrire all’utente particolari viste dei dati e, a richiesta, utilizzare queste sintesi come nuova chiave di accesso ai dati sui quali raffinare la ricerca con una più puntuale analisi dei contenuti testuali.

Un’ultima e importante prospettiva riguarda la migrazione all’interno di un’infrastruttura di ricerca europea come CLARIN<sup>10</sup> e DARIAH<sup>11</sup>, che costituiscono una risposta sicura al processo di recupero implementato, da un lato permettendo di valorizzare tale patrimonio testuale attraverso un’ampia diffusione indirizzata ad una platea più vasta, dall’altro offrendo al progetto un respiro internazionale.

Questa prospettiva si apre al più ampio panorama degli studi digitali nelle scienze umane e sociali, che si impegnano a preservare, documentare e rendere accessibili i dati, con l’utilizzo di standard di marcatura condivisi internazionalmente, ma anche a indicizzare i dati stessi in piattaforme comuni. È ormai divenuto un obiettivo irrinunciabile per ogni iniziativa di questo tipo affiancare al processo di standardizzazione finalizzato alla conservazione a lungo termine, l’interoperabilità e l’eventuale esposizione dei dati testuali nella ‘nuvola’ dei Linguistic Linked Open Data. Conclusa la fase di conversione, sarà quindi obiettivo del progetto descrivere adeguatamente la risorsa e integrarla nei cataloghi di CLARIN e/o DARIAH. In una prospettiva più lontana, ma comunque attuabile e tale da produrre una metadattazione compatibile e adeguata agli standard richiesti dalle infrastrutture, sarà possibile studiare come rendere la risorsa interrogabile attraverso la formulazione di ricerche “federate”, in grado di proiettare una singola ricerca sull’intera rete dell’infrastruttura.

## BIBLIOGRAFIA

[1] Becattini Iliaria, «Dalla Selva alla Cupola. Il trasporto del legname dell’Opera di Santa Maria del Fiore e il suo impiego nel cantiere brunelleschiano». *Gli anni della Cupola – Studi*. ISSN: 2364-6373 (2015): <http://duomo.mpiwg-berlin.mpg.de/STUDIES/study003/study003.html>

[2] Cappellotto Anna, «From codex to apps: the medieval manuscript in the age of its digital reproduction», (2020), DOI: <http://doi.org/10.6092/issn.2532-8816/11459>

[3] Coradeschi Francesco, Degl’Innocenti Emiliano, Di Meo Carmen, Sanesi Maurizio, Spadi Alessia, Spinelli Federica, «RESTORE: Opening data in Digital Humanities and Cultural Heritage». Pubblicato in: *RCDL 2022 18th Italian Research Conference on Digital Libraries Information and Research Science connecting to Digital and Library Science*, Ed. Università di Padova, Italy, (2022) pp. 24-25.

[4] Ciotti, Fabio, «Testo rappresentazione e computer. Contributi per una teoria della codifica informatica dei testi, in *Internet e le muse*». *La rivoluzione digitale nella cultura umanistica*, a cura di P. Nerozzi Bellman, Milano (1997), pp. 226-232

[5] Haines Margaret, «Gli anni della cupola. Una banca dati testuale della documentazione dell’Opera del Duomo di Firenze». *Bollettino del Centro di Ricerche Informatiche per i Beni Culturali* 7 (1997): 95-140; Eadem, «Gli anni della cupola. Archivio digitale delle fonti dell’Opera di Santa Maria del Fiore. Edizione di testi con indici analitici e strutturali». *Reti Medievali. Rivista* 3 (2002). <http://www.rmojs.unina.it/index.php/rm/article/view/264>

[6] Rosso Nadia, «Vertepe: una proposta di annotazione semantica XML-TEI della traduzione antica», *FuturoClassico* 6 (2020) 136-152. DOI: <https://doi.org/10.15162/2465-0951/1170>

<sup>9</sup> <http://dbtvm1.ilc.cnr.it/DBTOpa/>

<sup>10</sup> In particolare, CLARIN-IT è il nodo italiano di CLARIN (Common Language Resources and Technology Infrastructure).

<sup>11</sup> Nello specifico, DARIAH-IT è il nodo Italiano di DARIAH-EU (Digital Research Infrastructure for the Arts and Humanities).

- [7] Salvatori Enrica, Rosselli Del Turco R., Alzetta C., Di Pietro C., Mannari C., Miaschi A., «Il Codice Pelavicino tra edizione digitale e Public History». *Umanistica Digitale 1* (2017): 105-117. DOI: <https://doi.org/10.6092/issn.2532-8816/7232>
- [8] Sassolini Eva, Biffi Marco, «Strategie e metodi per il recupero di dizionari storici», *Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD)*, 15 – 17 gennaio 2020, Università del Sacro Cuore, Milano, ISBN: 978-88-942535-4-2, (2020), pp 235 – 239, DOI: <https://doi.org/10.6092/unibo/amsacta/6316>
- [9] Sassolini Eva, Cucurullo Sebastiana, Cinini Alessandra, «I corpora digitali: dall'obsolescenza tecnologica, alla salvaguardia e alla condivisione», *Conferenza GARR Selected papers*, Associazione Consortium GARR, Roma, maggio 2017, ISBN 978-88-905077-6-2, (2017) pp 31-35, Consortium GARR, Roma.
- [10] Sassolini Eva, Sassi Manuela, Cucurullo Sebastiana, Cinini Alessandra, Sbrulli Stefano, «Industrial Philology: problems and techniques of data and archives preservation for future generations», In: *GL 15 - The grey audit: a field assessment in grey literature : Fifteenth International Conference on Grey Literature ; Slovak Centre of Scientific and Technical Information, Bratislava, 2 - 3 December 2013 ; GL 15 ; conference proceedings (Bratislava, 2-3 december 2013)*. Abstract, (2013), pp. 73 - 77. D.J. Farace and J. Frantzen (eds.). TextRelease.
- [11] TEI Consortium. eds. "10 Manuscript Description." *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [4.5.0]. [2022-10-25]. TEI Consortium. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSFLT> (10/01/23).



# Ianus 2.0. Corrottele filologiche, congetture digitali

Giuseppe Ferrara<sup>1</sup>

<sup>1</sup> Università di Siena, Italia – giuseppe.ferrara@student.unisi.it

## ABSTRACT

Durante la redazione di un'edizione critica di un'opera antica, può accadere che il filologo classico si imbatta in un testo il cui significato gli appaia difficile da ricostruire oppure irrimediabilmente corrotto. In questo contributo, viene presentato *Ianus 2.0*, un *software* che nasce con lo scopo di aiutare il filologo classico a elaborare tutte le possibili congetture per una o più parole errate presenti all'interno di un testo greco o latino, contenente errori di maiuscola, minuscola o pronuncia. Tali congetture dovranno, poi, essere sottoposte al vaglio del filologo stesso, che potrà valutare, in prima persona, quale sia quella migliore per un determinato contesto. *Ianus 2.0* è un *software* scritto in linguaggio Python e si basa su un algoritmo di tipo deterministico, che, sfruttando un apposito *database*, è in grado di migliorare la propria velocità di risposta attraverso l'esperienza condotta sui testi immessi. L'efficacia di *Ianus 2.0* nell'elaborazione delle congetture rispetto ai filologi classici è stata valutata attraverso uno studio monocentrico, controllato e non randomizzato, che ha evidenziato come il *software* sia in grado di rintracciare la congettura attesa nel 100% dei casi, 95% CI [0.95, 1], contro un 25% dei filologi classici, 95% CI [0.17, 0.35], e di proporla, come prima congettura, nel 66% dei casi, 95% CI [0.55, 0.76], contro un 15% dei filologi classici, 95% CI [0.09, 0.24]. Un *software* con caratteristiche analoghe a quelle che vengono presentate in questo contributo potrebbe aprire la strada a nuovi filoni di ricerca nell'ambito della filologia digitale e delle *Digital Humanities*, che potrebbero giovare agli studi classici.

## PAROLE CHIAVE

Software di filologia digitale; Python; *emendatio*; errori paleografici; errori di pronuncia

## 1. LE FONDAMENTA DEL PROGETTO

Il presente contributo si pone l'obiettivo di offrire una presentazione sintetica di un progetto di filologia digitale sviluppato durante un periodo di ricerca di circa due anni, da me condotto in vista della stesura della mia tesi di laurea triennale, prima, e magistrale, poi. Il cuore di questo progetto è rappresentato da *Ianus 2.0*, un *software* che è stato implementato con lo scopo di aiutare il filologo classico a elaborare congetture per quei testi, greci e latini, che presentino, in uno o in più loro punti, corrottele caratterizzate da un livello di complessità variabile e riconducibili a tre categorie di errori: gli errori causati dall'utilizzo di scritture maiuscole nell'antigrafo; gli errori causati dall'utilizzo di scritture minuscole nell'antigrafo; gli errori nati dall'evoluzione della pronuncia delle lingue antiche tra la Tarda Antichità e il Medioevo. *Ianus 2.0* rappresenta l'aggiornamento di *Bentley 1.0*, un *software* da me precedentemente implementato per l'emendazione dei testi latini contenenti errori di minuscola. Il campo di utilizzo della nuova versione del *software* che presento di seguito è stato esteso rispetto a quello del suo predecessore, e potrebbe essere ulteriormente espanso nel prossimo futuro.

La necessità dello sviluppo di un *software* con caratteristiche quali quelle di *Ianus 2.0* nasce dal fatto che, nel campo di ricerca in cui esso si inserisce (ovvero, quello della filologia digitale e, più in generale, delle *Digital Humanities*), per quanto ne so, non è ancora stato sviluppato uno strumento che sia in grado di semplificare il processo di *emendatio* dei testi antichi. I soli strumenti che sono in grado di svolgere un lavoro simile, non di correzione ma di integrazione di testi greci, sono stati messi a punto nell'ambito del progetto di ricerca *DeepMind* di Google e si basano su algoritmi non deterministici di Intelligenza Artificiale (IA): si tratta di *Pythia* e *Ithaca*. Il primo è stato sviluppato con l'obiettivo di agevolare la ricostruzione dei caratteri mancanti all'interno delle epigrafi greche pervenute mutili e presenta un tasso di errore del 30,1%, contro un tasso di errore del 57,3% degli epigrafisti (Assael, Sommerschild, e Prag 2019). Il secondo nasce da una evoluzione del primo, a cui sono state aggiunte nuove funzionalità: rispetto a *Pythia*, infatti, *Ithaca* è in grado non soltanto di integrare le epigrafi greche pervenute mutili con un grado di accuratezza del 62%, ma anche di avanzare delle ipotesi sul loro luogo di produzione con un'accuratezza del 71% e di fornire una loro datazione con uno scarto di meno di 30 anni rispetto ai reali intervalli di datazione proposti dagli studiosi (Assael et al. 2022). Tali strumenti rappresentano il punto di partenza di questo progetto, che, tuttavia, adotta strategie implementative diverse da quelle di *Pythia* e *Ithaca*.

## 2. LA STRUTTURA DEL DATABASE

Per svolgere il compito per cui *Ianus 2.0* è stato progettato, è indispensabile che esso sia in grado di riconoscere se una parola greca o latina sia grammaticalmente valida. Il modo più semplice per verificare questa condizione è controllare se essa rientri nell'insieme dei termini greci e latini conosciuti fino ad oggi. Per implementare questa funzione, dunque, è stato fornito al *software* un archivio contenente le parole greche e latine testimoniate all'interno dei documenti antichi che sono giunti fino ai nostri giorni. Questo archivio è stato organizzato nella forma di un *database* relazionale dotato di una struttura specifica, che si può vedere sintetizzata, in figura 1, attraverso un diagramma entità-relazione (E-R).



Figura 10. Diagramma E-R in notazione Crow's Foot del database di *Ianus 2.0*. Elaborazione realizzata tramite il servizio online [dbdiagram.io](http://dbdiagram.io).

Il *database* di *Ianus 2.0* è strutturato in quattro tabelle: *PAROLE\_GRECHE*, *PAROLE\_LATINE*, *CORREZIONI\_GRECHE* e *CORREZIONI\_LATINE*. Le prime due assolvono la funzione di accogliere le parole greche e latine testimoniate all'interno della documentazione posseduta in entrambe le lingue. La tabella *PAROLE\_GRECHE* presenta due attributi di tipo *text*, identificati, rispettivamente, tramite le etichette *parola* e *parola\_senza\_spiriti\_accenti*: il primo attributo, che funge anche da *primary key*, ospita tutte le parole greche esistenti, traslitterate in caratteri latini; il secondo attributo accoglie le traslitterazioni senza spiriti e accenti. L'importanza di questo secondo campo deriva dal fatto che, quando *Ianus 2.0* processa le parole da correggere, lavora su caratteri alfabetici puri, ma, alla fine del processo di correzione, deve essere in grado di risalire alla parola originaria dotata di spiriti e accenti per mostrarla all'utente. Inoltre, poiché, in molti casi, gli spiriti e gli accenti sono anche elementi decisivi per poter distinguere, in greco, una parola da un'altra, a una stessa traslitterazione priva di questi due elementi possono essere associate più parole distinte: per questo motivo, la relazione che sussiste tra le due colonne *parola\_senza\_spiriti\_accenti* e *parola* è del tipo *uno-a-molti*. La tabella *PAROLE\_LATINE*, invece, presenta un unico attributo di tipo *text*, identificato tramite l'etichetta *parola*, che è stato pensato per accogliere, come *record*, tutte le parole latine esistenti.

Le tabelle *CORREZIONI\_GRECHE* e *CORREZIONI\_LATINE* presentano una struttura sostanzialmente speculare, articolata in sei attributi, di cui il primo e l'ultimo di tipo *integer* e gli altri di tipo *text*: *id*, *parola\_iniziale*, *correzione*, *parola*, *tipologia\_errore* e *livello\_analisi*. L'attributo *id* è auto-incrementale ed è stato introdotto con lo scopo di assolvere la funzione di *primary key*. L'attributo *parola\_iniziale* contiene la parola immessa dall'utente all'inizio del processo di elaborazione delle congetture e serve per tracciare il punto da cui inizia tale processo. L'attributo *correzione* è stato pensato per accogliere le singole congetture che vengono calcolate dal *software* durante la correzione del testo immesso. L'attributo *parola* ospita la variante della parola iniziale su cui il *software* lavora per elaborare la correzione medesima. L'attributo *tipologia\_errore* è stato pensato per memorizzare, all'interno del *database*, la tipologia e il numero di errori che sono stati riscontrati nella parola iniziale e che hanno permesso di elaborare la congettura finale. Questo dato viene memorizzato sotto forma di lista Python a sei *item*, divisibili in tre coppie: in ognuna di esse, il primo elemento identifica il numero di errori, il secondo, che può avere tre valori distinti (1 per gli errori di minuscola, 2 per gli errori di maiuscola e 3 per gli errori di pronuncia), la tipologia di errore. L'attributo *livello\_analisi* accoglie il numero di livelli di analisi selezionato dall'utente prima dell'inizio del processo di correzione: questo dato è stato introdotto per indicare la profondità dell'analisi operata dal *software* durante il processo di correzione delle parole errate immesse e per avere una misura della complessità degli errori riscontrabili in ciascuna parola immessa. In breve, il concetto di *livello di analisi* coincide con il numero massimo di volte che il *software* può agire, modificandola, su ciascuna lettera che compone ciascuna delle parole del testo sottoposto alla sua analisi: per esempio, i testi con un livello di analisi potranno essere corretti modificando *al massimo una volta*, in maniera alternativa ed esclusiva, ognuna delle lettere che costituiscono ciascuna delle loro parole. Tra i campi *parola\_iniziale*, *correzione* e *parola* della tabella *CORREZIONI\_GRECHE* e il campo *parola* della tabella *PAROLE\_GRECHE* esistono tre relazioni del tipo *uno-a-molti*. Tra i campi *parola\_iniziale*, *correzione* e *parola* della

tabella *CORREZIONI\_LATINE* e il campo *parola* della tabella *PAROLE\_LATINE* esistono tre relazioni del tipo *uno-a-uno*.

Per popolare i campi delle tabelle *PAROLE\_GRECHE* e *PAROLE\_LATINE*, è stata sfruttata una risorsa libera messa a disposizione dei ricercatori dal progetto *Perseus Digital Library*: si tratta dell'archivio, in formato *.tar.gz*, con tutti i *file* del progetto (balmas, un e\_gweaver, e perseus\_rsingh 2016). In particolare, all'interno di questo archivio sono presenti due *file* (*greek.words* e *latin.words*) che contengono un elenco completo delle parole greche e latine che compaiono nelle risorse disponibili all'interno del progetto. I dati contenuti in questi *file* sono stati caricati, con alcune debite modifiche finalizzate ad eliminarne la ridondanza e gli errori di trascrizione, nel *database* di *Ianus 2.0*, mantenendo, però, le regole di traslitterazione delle parole greche adottate in *Perseus Digital Library*.

### 3. L'ALGORITMO DI CORREZIONE

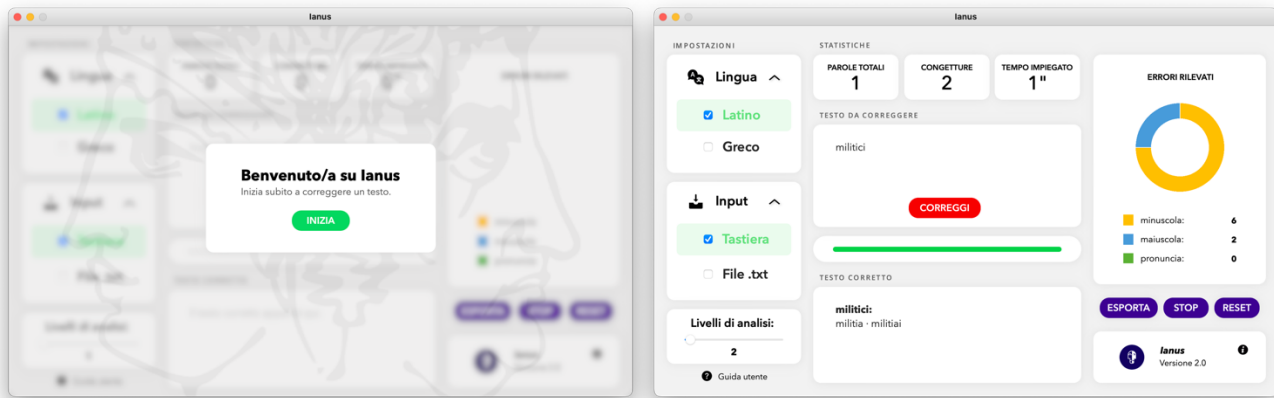


Figura 11. Schermate visualizzate durante una sessione di esecuzione di *Ianus 2.0*.

La natura meccanica e regolare degli errori di cui mi sono occupato mi ha indotto a scegliere di implementare un algoritmo di tipo deterministico, in cui, a partire dai medesimi dati di *input*, si ottengono sempre i medesimi dati di *output*, in quanto il percorso segnato dalle sue istruzioni risulta obbligato (Peres 2020). Di tale algoritmo, in questo paragrafo, voglio presentare i passaggi fondamentali. Durante una sessione di esecuzione del *software*, che si può vedere rappresentata in figura 2, si richiede all'utente di inserire un testo che contenga una o più parole errate: quest'ultimo può essere, alternativamente, immesso da tastiera oppure importato da un file di testo con estensione *.txt*. Dopo che l'utente clicca sul tasto *CORREGGI*, il *software* inizia a dividere il testo inserito in linee e opera su ciascuna di esse isolatamente, in modo tale da prevenire un sovraccarico della memoria RAM del dispositivo, dovuto al caricamento dell'intero testo su di essa; ogni linea viene, poi, divisa in una lista contenente tutte le parole che la costituiscono. Per ciascuna parola inclusa in questa lista, *Ianus 2.0* verifica se siano già presenti, all'interno del *database*, delle sue congetture valide e se la parola in questione sia già presente nella struttura dati di tipo *set* in cui vengono inserite le parole per cui, in quella sessione di esecuzione del *software*, sono già state calcolate le congetture. Se entrambi questi *test* hanno esito negativo, la parola analizzata viene aggiunta a una coda, ovvero a una struttura dati che adotta una politica di tipo *FIFO* (Cormen, Leiserson, e Rivest 1994, 189-192), contenente le parole che devono essere corrette; altrimenti, si estraggono le congetture già presenti nel *database* per quella parola e le si mostrano alla fine del processo di correzione. Fintantoché la coda presenta elementi al suo interno, *Ianus 2.0* estrae da essa la parola da correggere, la inserisce all'interno del *set* contenente le parole per cui sono state generate congetture durante quella sessione di esecuzione e spacchetta, isolandoli, i caratteri della parola da correggere, in modo tale da poter operare, singolarmente, su ognuno di essi. Per ciascun carattere *x* presente all'interno di questa lista viene applicata la funzione di correzione, che opera come segue. Se il carattere *x* coincide con un carattere che può essere scambiato con un altro per un errore di maiuscola, minuscola o pronuncia, allora il carattere *x* viene sostituito con quello corretto. Se la parola ottenuta tramite questa sostituzione è una parola esistente nella lingua di riferimento, allora essa viene considerata dal *software* come possibile congettura e viene aggiunta a una lista, che verrà mostrata all'utente alla fine del processo di correzione, contenente la parola di partenza e le sue congetture e i dati prodotti vengono salvati all'interno del *database*.

Quest'ultimo passaggio risulta di fondamentale importanza in quanto consente il miglioramento delle prestazioni dell'algoritmo del *software* e l'incremento della sua velocità di risposta a seguito dell'esperienza acquisita su un *set* di dati durante un periodo di addestramento. Salvando, nel *database*, le congetture già elaborate per determinate parole e le parole

errate di cui queste ultime sono possibili correzioni, *Ianus 2.0* è in grado di recuperare, senza elaborarle nuovamente, le congetture già calcolate per una determinata parola corrotta, nel caso in cui quest'ultima si ripresenti in un altro punto dello stesso testo o in un altro testo diverso. Una conseguenza di questo modo di procedere del *software* è anche che, quanto più esso fa esperienza sui testi che gli vengono sottoposti, tanto più la sua velocità di risposta aumenta e i tempi necessari per portare a termine il processo di correzione si riducono.

#### 4. UNO STUDIO MONOCENTRICO, CONTROLLATO, NON RANDOMIZZATO

Quando si sviluppa un nuovo strumento per risolvere un problema noto, l'adozione di un approccio di tipo scientifico richiede che venga effettuata una valutazione dell'efficacia di quest'ultimo, prima che il suo uso possa essere generalizzato. Poiché questo lavoro è ispirato ai principi delle scienze dure, ho voluto fare mia questa impostazione metodologica e ideare uno studio con cui testare il *software* da me sviluppato.

L'ipotesi alla base dello studio che ho condotto è stata che *Ianus 2.0* potesse fornire un contributo effettivo alla correzione degli errori di maiuscola, minuscola e pronuncia presenti nei testi antichi, sgravando il filologo classico di una parte considerevole del suo lavoro e consentendo a quest'ultimo di esaminare congetture possibili, ma non ancora esplorate. L'obiettivo della sperimentazione, dunque, è coinciso con la valutazione della maggiore o minore efficacia di *Ianus 2.0* rispetto a un filologo umano nel correggere tali categorie di errori. L'efficacia dell'intervento (di *Ianus 2.0* o del filologo umano) è stata valutata nei termini della capacità di trovare la parola corretta (considerata il *gold standard* dell'esperimento) all'origine di una data corruzione. In sede di analisi dei dati, è stata valutata anche la probabilità che la parola corretta comparisse tra le prime tre congetture elaborate. Questa rilevazione ha permesso di fornire, all'utente finale del *software*, alcune utili indicazioni operative per sfruttare al meglio le correzioni proposte dal *software* stesso.

Per raggiungere l'obiettivo dichiarato, si è progettato uno studio sperimentale monocentrico, controllato e non randomizzato. La sperimentazione è stata condotta presso il Dipartimento di Filologia e critica delle letterature antiche e moderne (DFCLAM) dell'Università degli Studi di Siena, tra il 29 marzo 2023 e l'8 aprile 2023, dopo aver reclutato due filologi classici, l'uno dotato di una formazione specialistica nella lingua greca, l'altra dotata di una formazione specialistica nella lingua latina, considerati come trattamenti alternativi al *software* nel gruppo di controllo.

La sperimentazione ha previsto lo svolgimento, in parallelo, di due sessioni di correzione di 80 parole errate (40 greche e 40 latine), che hanno formato il campione dello studio per la popolazione di riferimento rappresentata dall'insieme di tutte le parole greche e latine che possono essere interessate da errori di maiuscola, minuscola e pronuncia. Per ottenere il campione di parole errate utilizzato, si sono selezionate, in maniera casuale, le parole originarie corrette da una serie di testi greci e latini presenti nel *corpus* di *Perseus Digital Library* (Crane s.d.), sulla base dei seguenti criteri di inclusione: 1) presenza di lettere o gruppi di lettere suscettibili degli errori di maiuscola, minuscola e pronuncia; 2) numero di caratteri variabile. Successivamente, le parole del campione sono state suddivise in due liste da 20 *item* ciascuna per il greco e per il latino, al fine di creare due diversi contesti di correzione con livelli di complessità variabili per ciascuna lingua (il primo di difficoltà minore, il secondo di difficoltà maggiore), in modo tale da verificare l'efficacia del *software* e dei filologi umani in entrambi. Le parole delle liste così ottenute sono state alterate inserendovi, manualmente, alcuni degli scambi noti per le categorie di errori considerate, secondo i seguenti criteri: 1) la prima lista è stata manipolata in modo tale da contenere soltanto parole in cui fosse presente un solo errore di una sola categoria; 2) la seconda lista è stata manipolata in modo tale da contenere soltanto parole in cui fossero presenti due errori della stessa categoria o di categorie diverse, che interessassero una stessa lettera in due stadi successivi di alterazione oppure due lettere diverse in un solo stadio di alterazione.

Le due liste di parole greche e le due liste di parole latine sono state assegnate sia al *software* sia al filologo competente. Il disegno sperimentale impiegato è un disegno a due gruppi: il gruppo sperimentale, composto dalle liste di parole assegnate al *software*, e il gruppo di controllo, composto dalle liste di parole assegnate al filologo competente. Al *software* il diverso grado di complessità delle parole presenti nelle tre liste è stato segnalato indicando, all'inizio del processo di correzione, il *livello di analisi* da impiegare per ciascuna lista di parole (uno per la prima lista, due per la seconda). Ai filologi il diverso grado di complessità delle parole presenti nelle tre liste è stato segnalato indicando il numero massimo di errori che potevano essere ipotizzati per ciascun gruppo di parole (uno per la prima lista, due per la seconda). Nel corso dell'esperimento, inoltre, è stato chiesto ai filologi di scrivere le congetture da loro proposte per ciascuna parola su un foglio *Excel* in un ordine di probabilità decrescente, in cui la prima congettura indicata per la parola errata coincidesse con quella ritenuta più probabile, l'ultima con quella ritenuta meno probabile.

Per valutare l'efficacia del *software* rispetto ai filologi umani, è stata considerata la variabile quantitativa *punteggio totalizzato*, calcolata per ciascuna parola (greca e latina) di ciascuna lista. Il valore di tale variabile è stato stabilito assegnando un punteggio, al *software* e ai filologi, per ciascuna parola corretta originaria recuperata, sulla base di un criterio posizionale: si è attribuito un punteggio pari a 3, quando la parola corretta originaria compariva come prima congettura tra

quelle formulate, un punteggio pari a 2, quando la parola corretta originaria compariva come seconda congettura, un punteggio pari a 1, quando la parola corretta originaria compariva come terza congettura, e un punteggio pari a 0 in tutti gli altri casi. La distribuzione dei punteggi assegnati a *Ianus 2.0* e ai filologi classici è rappresentata in figura 3.

Per ciascuna lista di parole greche e latine, l'ipotesi nulla  $H_0$  da sottoporre a *test* è stata che la variabile rappresentata dalla differenza fra il punteggio totalizzato dal *software* e quello totalizzato dai filologi per ciascuna parola analizzata assumesse un valore medio uguale a 0. L'analisi a due code ha utilizzato, come soglia di significatività, un *p-value* di 0.05 e la verifica delle ipotesi è stata condotta mediante il *test* non parametrico di Wilcoxon per dati appaiati. Per il calcolo degli intervalli di confidenza, invece, è stato utilizzato lo *Score Method* di Wilson. Per l'analisi statistica, infine, è stato utilizzato *IBM SPSS Statistics*.

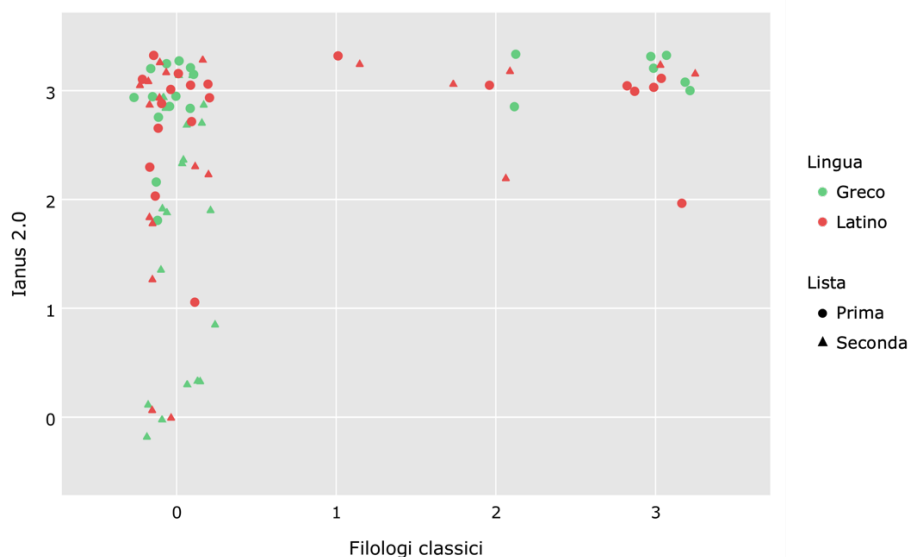


Figura 12. Distribuzione dei punteggi assegnati a *Ianus 2.0* e ai filologi classici per lingua e per lista.

Lista	<i>Ianus 2.0</i>		Filologo/a	
	%	95% CI	%	95% CI
Prima (greco)	100	[0.84, 1]	35	[0.18, 0.57]
Seconda (greco)	70	[0.48, 0.85]	0	[0, 0.16]
Prima (latino)	100	[0.84, 1]	35	[0.18, 0.57]
Seconda (latino)	90	[0.70, 0.97]	30	[0.15, 0.52]
<b>TOTALE</b>	<b>90</b>	<b>[0.81, 0.95]</b>	<b>25</b>	<b>[0.17, 0.35]</b>

Tabella 1. Percentuali dei casi in cui *Ianus 2.0* e i filologi classici hanno rintracciato la parola corretta originaria nelle prime tre posizioni, con intervalli di confidenza (CI) arrotondati alla seconda cifra decimale.

Lista	<i>Ianus 2.0</i>		Filologo/a	
	%	95% CI	%	95% CI
Prima (greco)	90	[0.70, 0.97]	25	[0.11, 0.47]
Seconda (greco)	35	[0.18, 0.57]	0	[0, 0.16]
Prima (latino)	80	[0.58, 0.92]	25	[0.11, 0.47]
Seconda (latino)	60	[0.39, 0.78]	10	[0.03, 0.3]
<b>TOTALE</b>	<b>66</b>	<b>[0.55, 0.76]</b>	<b>15</b>	<b>[0.09, 0.24]</b>

Tabella 2. Percentuali dei casi in cui *Ianus 2.0* e i filologi classici hanno rintracciato la parola corretta originaria in prima posizione, con intervalli di confidenza (CI) arrotondati alla seconda cifra decimale.

L'analisi dei dati raccolti ha mostrato che *Ianus 2.0* è stato in grado di rintracciare la parola corretta originaria nel 100% dei casi ad esso sottoposti, 95% CI [0.95, 1], contro un 25% dei filologi classici, 95% CI [0.17, 0.35]. Se si considerano soltanto le prime tre posizioni tra le congetture proposte dal *software* e dai filologi, la percentuale di successo di *Ianus 2.0* si attesta al 90%, 95% CI [0.81, 0.95], mentre quella dei filologi classici rimane invariata. Le percentuali dettagliate dei successi del *software* e dei filologi, con i rispettivi intervalli di confidenza, sono riportate nelle tabelle 1-2.

L'applicazione del test di Wilcoxon per dati appaiati alle coppie di punteggi (punteggio di *Ianus 2.0* e punteggio del filologo) relative a ogni lista ha permesso di rigettare, in tutti i casi, l'ipotesi nulla  $H_0$ , evidenziando una differenza statisticamente significativa ( $p < 0.001$ ) tra il *software* e il filologo competente. In media, inoltre, la differenza tra i punteggi di *Ianus 2.0* e di ciascun filologo è stata positiva: il *software* ha, cioè, ottenuto punteggi mediamente più alti rispetto ai filologi umani.

Lo studio ha evidenziato che, nella maggioranza dei casi, l'algoritmo di correzione di *Ianus 2.0* è stato in grado di presentare la correzione attesa come prima congettura tra quelle elaborate. La variazione osservata nel comportamento del *software* sembra essere collegata al diverso livello di complessità delle parole sottoposte alla sua analisi e alla lingua delle parole immesse: più è alto il livello di complessità, più il *software* tende a mostrare la congettura attesa in una posizione distante da quella iniziale, e ciò accade, in particolare, nel caso in cui vengano analizzate parole greche. Questi risultati permettono di affermare che, nel caso in cui l'utilizzatore finale del *software* voglia trovare una congettura per una data parola immessa, prima di passare in rassegna tutte le correzioni elaborate dal *software* stesso, può verificare che una tra le prime tre correzioni restituisca un senso al testo e può limitarsi a esaminare le congetture restanti solo nel caso in cui le prime tre risultino non soddisfacenti e solo nel caso in cui supponga la presenza, nella parola analizzata, di più di un errore paleografico o di pronuncia. La valutazione delle congetture proposte oltre la terza è consigliabile soprattutto nel caso di una parola greca in cui si ipotizzino corrotte multiple.

Lo studio ha evidenziato anche che *Ianus 2.0* è stato in grado di rintracciare la parola corretta originaria estrapolata dal contesto in tutti i casi ad esso sottoposti, mentre i filologi umani sono stati in grado di fare ciò soltanto in un quarto dei casi da loro analizzati. Tale risultato nei gruppi trattati dai filologi umani deve essere stato influenzato da due fattori: dal livello di complessità della lista di parole sottoposta alla loro analisi e dall'assenza di un contesto che agevolasse il processo di *emendatio*. In merito al primo punto, si può dire che, con l'aumentare del livello di complessità delle parole analizzate, si è osservato un peggioramento della *performance* di entrambi i filologi, meno marcata per il latino e più marcata per il greco. In merito al secondo punto, sarebbe interessante effettuare, in futuro, uno studio che valuti come il contesto influenzi l'elaborazione delle congetture da parte dei filologi umani e che determini la validità di un approccio algoritmico che tenga in considerazione anche questo aspetto.

## 5. CONCLUSIONI

Il progetto di cui si è parlato nel corso di questo lavoro rappresenta una nuova frontiera di ricerca nell'ambito della filologia digitale e delle *Digital Humanities*. Sebbene, ad oggi, le categorie di errori di cui esso si occupa siano limitate, in futuro è possibile che esse vengano estese, sfruttando la combinazione di algoritmi di Intelligenza Artificiale e di algoritmi di tipo deterministico, fino a coprire l'intero spettro degli errori riscontrabili nei testi antichi. Poiché mi sono trovato ad operare in autonomia e in assenza di finanziamenti, ho scelto di adottare soluzioni esecutive il più possibile semplici e lineari, ma comunque adeguate agli obiettivi che mi ero prefissato. Per l'implementazione, in tempi brevi, di nuove funzionalità complesse attraverso l'utilizzo di principi mutuati dagli ambiti dell'Intelligenza Artificiale e del *Machine Learning* sarebbe necessario un lavoro di équipe, simile a quello realizzato per lo sviluppo di *Pythia* e *Ithaca*. In ogni caso, la versione che è stata presentata in questo contributo potrebbe rappresentare il punto di partenza per lo sviluppo di un *software* che sia in grado di creare, autonomamente, edizioni digitali delle opere classiche, greche e latine, passando al vaglio, in poco tempo, migliaia di lezioni e di congetture e proponendo, sulla base di indizi contestuali e di raffronti con altre opere degli stessi autori, le lezioni e le varianti ritenute più verosimili. Per poter raggiungere un simile risultato, il lavoro da compiere è ancora molto. Nel caso in cui, però, uno strumento di questo tipo potesse essere realizzato, i suoi vantaggi potrebbero essere determinanti nell'ambito degli studi classici.

In merito allo studio di cui si sono riportati i risultati, sebbene esso si sia rivelato sufficientemente potente da offrire dati solidi per rispondere agli obiettivi per cui era stato progettato, presenta alcuni limiti importanti che dovranno essere presi in considerazione per ricerche future. In primo luogo, data la natura preliminare della sperimentazione, è stato possibile reclutare soltanto due filologi per controllare i risultati del *software*. Poiché la capacità di elaborare congetture, pur presentando alcuni elementi standardizzati, dipende dalla formazione del singolo filologo, sarebbe interessante effettuare, in futuro, il medesimo esperimento reclutando diversi specialisti del settore. In secondo luogo, il campione di parole greche e latine scelto, per quanto rappresentativo della popolazione di riferimento, è risultato piuttosto limitato e ha permesso di

ottenere intervalli di confidenza, in molti casi, ampi per le *performance* del *software* e dei filologi. Sarebbe, dunque, interessante valutare, in futuro, il comportamento di questi ultimi su un numero più corposo di parole per ottenere risultati più precisi.

## BIBLIOGRAFIA

- [1] Assael, Yannis, Thea Sommerschild, e Jonathan Prag. «Restoring ancient text using deep learning: a case study on Greek epigraphy». arXiv, 14 ottobre 2019. <http://arxiv.org/abs/1910.06262>
- [2] Assael, Yannis, Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, e Nando de Freitas. «Restoring and Attributing Ancient Texts Using Deep Neural Networks». *Nature* 603, fasc. 7900 (10 marzo 2022): 280–83. <https://doi.org/10.1038/s41586-022-04448-z>
- [3] balmas, perseus\_gweaver, e perseus\_rsingh. «Perseus' Java Hopper download | SourceForge.net», 26 gennaio 2016. <https://bit.ly/3Z17fPw>
- [4] Bland, Martin. *Statistica medica*. Seconda edizione. Santarcangelo di Romagna: Maggioli Editore, 2019.
- [5] Braccini, Tommaso. *La scienza dei testi antichi. Introduzione alla filologia classica*. Firenze: Le Monnier Università, 2017.
- [6] Brody, Tom. *Clinical trials: study design, endpoints and biomarkers, drug safety, and FDA and ICH guidelines*. Second edition. Amsterdam; Boston: Elsevier/AP, 2016.
- [7] Buttu, Marco. *Programmare con Python. Guida completa*. Milano: LSWR, 2014.
- [8] Cormen, H. Thomas, Charles E. Leiserson, e L. Ronald Rivest. *Introduzione agli algoritmi*. Milano: Jackson Libri, 1994.
- [9] Coronel, Carlos, e Steven Morris. *Database systems: design, implementation, and management*. 13th e. Australia; United States: Cengage Learning, 2019.
- [10] Crane, R. Gregory. «Greek and Roman Materials». Consultato 2 aprile 2023. <https://bit.ly/3G8TeTz>
- [11] Delz, Joseph. «Critica testuale ed ecdotica». In *Introduzione alla filologia latina*, a cura di Fritz Graf. Roma: Salerno Editrice, 2003.
- [12] Fränkel, Hermann. *Testo critico e critica del testo*. Firenze: Le Monnier, 1983.
- [13] Havet, Louis. *Manuel de critique verbale appliquée aux textes latins*. Paris: Hachette, 1991.
- [14] Holistics Software. «dbdiagram.io - Database Relationship Diagrams Design Tool». Consultato 4 marzo 2023. <https://bit.ly/3JcyNa2>
- [15] Lutz, Mark. *Imparare Python*. Milano: Hops tecniche nuove, 2011.
- [16] Peres, Ennio. *Che cosa sono gli algoritmi*. Milano: Salani, 2020.
- [17] Python Software Foundation. «sqlite3 — DB-API 2.0 interface for SQLite databases — Python 3.11.2 documentation», 2023 2001. <https://bit.ly/3ZcXKaV>
- [18] Reynolds, D. Leighton, e G. Nigel Wilson. *Copisti e filologi. La tradizione dei classici dall'antichità ai tempi moderni*. Roma; Padova: Antenore, 2016.
- [19] Riverbank Computing Limited. «Introduction — PyQt Documentation v6.4.1», 2023. <https://bit.ly/3SRb5U6>
- [20] Shamley, Delva, e Brenda Wright, a c. di. *A comprehensive and practical guide to clinical trials*. London, United Kingdom; San Deigo, CA, United States: Elsevier/AP, Academic Press, an imprint of Elsevier, 2017.
- [21] West, L. Martin. *Critica del testo e tecnica dell'edizione critica*. Palermo: L'Epos, 1991.
- [22] Willis, James. *Latin Textual Criticism*. Urbana; Chicago; London: University of Illinois Press, 1972



# Verso una «volontà d'archivio» digitale

Lucia Giagnolini<sup>1</sup>

<sup>1</sup> Università di Bologna, Italia – lucia.giagnolini2@unibo.it

## ABSTRACT

La "volontà d'archivio" è un concetto che si riferisce alla volontà di conservare e tramandare il proprio archivio. Nel mondo analogico si possono ricercare tracce di "volontà d'archivio" nelle modalità in cui i documenti sono stati organizzati, conservati e postillati dall'autore. Se per il cartaceo l'intenzione conservativa è stata per secoli agevolata dalla stabilità del supporto scrittorio, il digitale ha modificato drasticamente il rapporto dell'autore con la produzione e la gestione documentaria. Dall'analisi di interviste e sondaggi, emerge come la maggior parte degli scrittori trascurino o diano poca importanza ai propri archivi digitali, in un'epoca in cui, al contrario, la "volontà d'archivio" risulta assolutamente cruciale per la preservazione futura. Per questo motivo, le istituzioni individuate per la conservazione del digitale d'autore devono assumersi più responsabilità rispetto al passato per salvaguardare questo patrimonio. Gli sforzi istituzionali dovrebbero concentrarsi, in particolare, su ricerca e sviluppo di tecniche informatiche, archivistiche e filologiche in grado di gestire lo spettro di complessità di questi archivi, dall'acquisizione alla consultazione, nel rispetto della normativa vigente.

## PAROLE CHIAVE

Archivi nativi digitali di persona; digitale d'autore; archivistica informatica

## 1. LA PROSPETTIVA ANALOGICA

Gli archivi degli scrittori conservano carte, volumi e materiali che rappresentano le dimensioni di elaborazione e ricerca alla base delle loro opere letterarie. Le modalità in cui questi documenti sono stati organizzati e conservati dall'autore ci parlano della sua consapevolezza riguardo all'importanza di conservare e tramandare le proprie carte personali. Per descrivere questo fenomeno, è stato individuato il concetto di "volontà d'archivio" [1], una nuance della volontà autoriale che testimonia la fiducia degli scrittori nella possibilità di offrire alla posterità un'immagine parallela e complementare delle loro opere. Nel recente convegno *Volontà d'archivio. L'autore, le carte, l'opera*, svoltosi a Padova nel settembre 2022, una serie di interventi ha messo in luce come questo tema possa emergere dagli archivi dei più noti autori della letteratura italiana. In occasione dell'iniziativa, esperti di carte d'autore hanno cercato di individuare i rapporti dialettici che intercorrono tra i documenti archivistici, le opere e le indicazioni degli autori in relazione alla disposizione verso la conservazione e la fruizione delle proprie carte [16]. Per sondare l'esistenza di una "volontà d'archivio", esplicita o implicita, occorre effettuare uno studio strutturale di documenti quali lasciti testamentari o inventari, così come un'«analisi ravvicinata delle note d'autore, le cosiddette "meta-postille", che mostrano, nella stratificazione diacronica degli interventi, le riflessioni dell'autore sul proprio archivio» [16]. In Italia, Petrarca è considerato il primo autore in cui è riconoscibile una "volontà d'archivio", grazie alla decisione di conservare, con il *Codice degli abbozzi*, non solo l'opera nella sua versione definitiva, ma anche nelle tracce della sua genesi, dando il via ad una "tradizione" di manoscritti d'autore unica nel panorama europeo dal punto di vista genetico [10][16]. La conservazione di questo eccezionale patrimonio archivistico si deve – oltre che alla cura di bibliotecari e archivisti nel corso dei secoli – al dato sostanziale della stabilità del supporto cartaceo di fronte allo scorrere del tempo. Ma come si esplicita, oggi, la "volontà d'autore", in un'epoca in cui il digitale ha modificato drasticamente il rapporto con il testo e il documento? Come deve comportarsi un autore che desidera conservare il suo archivio digitale?

## 2. VOLONTÀ D'ARCHIVIO NEL MONDO DIGITALE

Per esplorare la misura in cui la "volontà d'archivio" si presenta fra gli autori contemporanei, possiamo innanzi tutto ricercare riferimenti espliciti in articoli, dichiarazioni e interviste.

L'utilizzo del computer per la scrittura iniziò a diffondersi tra gli autori sin dai primi anni '80. Nel panorama italiano, le ricerche di Domenico Fiormonte sul rapporto nascente fra scrittura e word processor individuarono un primo nucleo di scrittori "entusiasti" del nuovo strumento, come Umberto Eco, Claudio Pozzoli, Carmen Covito e Luciano De Crescenzo, immersi in un contesto ancora intriso di scetticismo [12][13]. Umberto Eco, in particolare, fu un pioniere anche nel promuoverne l'introduzione sistematica della tecnologia negli ambienti umanistici e nella didattica [9]: è facile rintracciare in lui una delle prime voci in cui emerge una "volontà d'autore" anche per il digitale. A tal proposito, Emmanuela Carbè,

in *Digitale d'autore. Macchine, archivi, letterature* [5], riporta un'intervista di Jean-Claude Carrière a Umberto Eco, in cui alla domanda «la tua casa va a fuoco: sai quale opere cercheresti di proteggere?» Eco rispondeva: «dopo che ho parlato così bene dei libri, lasciami dire che io mi porterei via il mio disco rigido esterno di 250 giga, che contiene tutti i miei scritti degli ultimi trent'anni» [6].

Proprio dal Master in Editoria che Umberto Eco fondò nel 2001 presso l'Università di Bologna, è nato, venti anni dopo, il progetto *A carte scoperte* [15], un'inchiesta sulle modalità di lavoro di scrittrici e scrittori contemporanei. Il volume raccoglie interviste rivolte a ventidue scrittori italiani, ognuna delle quali si conclude con una domanda dedicata al loro archivio: «Ha mai pensato di destinare il suo archivio a qualche istituzione culturale? E se sì, a quale e perché?». Le risposte sono alquanto sorprendenti: in relazione all'ipotesi di una donazione d'archivio, quattro autori si esprimono per un categorico “no”; in sei ritengono di doverci ancora riflettere; in dodici si sono mostrati propositivi, se non entusiasti all'idea. Tuttavia, dall'analisi delle risposte positive – che in prima istanza sembrerebbero testimoniare un certo grado di volontà di conservazione – emerge ancora la tendenza di considerare “archivio” solamente la produzione cartacea o, pur nella consapevolezza della presenza di un archivio digitale, ad esso viene generalmente attribuita minor importanza. Disposizione che emerge chiaramente dalle parole di Paola Capriolo, ad esempio, che ricorda come «molti anni fa, quando ancora scrivevo a penna, ho donato le prime stesure dei miei libri già pubblicati al Fondo manoscritti istituito da Maria Corti presso l'Università di Pavia», ma «adesso, con la scrittura elettronica, mi domando se una donazione del genere avrebbe ancora lo stesso significato» [15]. Anche Marcello Fois sottolinea la natura ibrida del suo archivio, manifestando un pensiero molto simile a quello di Capriolo nella percezione di analogico e digitale: «In realtà ci ho pensato, considerato il fatto che io appartengo a quella generazione di autori che ha prodotto ancora molto cartaceo e che quindi ha a disposizione una quantità notevole di materiale tangibile da conservare. Certo, a questo punto ne esiste molto di più da conservare nella memoria del mio Mac, ma credo che sia più interessante poter toccare con mano le bozze reali con le annotazioni» [15]. La risposta fornita da Andrea Tarabbia è l'unica in cui la consapevolezza del valore digitale emerge chiaramente, anteposta – almeno sintatticamente – alla considerazione sul cartaceo: «collaboro da anni con Fondazione Mondadori, a Milano; conosco bene l'archivio e gli archivisti. Ogni tanto, scherzando, ci diciamo che prima di morire manderò loro una mail con tutte le mie password e che mia moglie saprà a chi destinare il Fondo Tarabbia quando non ne potrà più di tutta questa carta» [15]. Fra gli autori intervistati, Paolo di Paolo è il solo ad aver già conferito parte del suo archivio digitale al progetto Pavia Archivi Digitali nel 2015 [5], confluito poi nel patrimonio documentale del Centro Manoscritti dell'Università di Pavia. Se questo ci parla certamente dell'apertura dell'autore nei confronti del tema, è interessante sottolineare il punto di vista con cui effettuò la donazione: «spero soprattutto di poter donare, un giorno, la biblioteca. Ma l'archivio, non so, non credo sia così degno di interesse; e non lo dico per falsa modestia. Ho risposto comunque alla richiesta del Fondo Manoscritti dell'Università di Pavia di donare – più per ragioni sperimentali che di reale interesse dei materiali – un mio archivio digitale [...] Era più che altro un supporto all'impresa di ripensare la filologia d'autore nel ventunesimo secolo e di provare ad interrogare diversamente la fitta variantistica dei testi nati sui programmi di video scrittura».

Diversi studi hanno cercato di analizzare la concezione del valore dei propri scritti digitali da parte dei soggetti produttori [3] [19][20], senza la quale non è possibile immaginare la formazione di una “volontà d'archivio”. Marshall, Bly e Brun-Cottan collegano strettamente il valore assunto da un file per il suo autore alle modalità e alle precauzioni con cui viene gestito [19]; Becker e Nouges, pur condividendo la prospettiva, sottolineano come nell'equazione valore-gestione – almeno nel caso degli scrittori – occorra tenere in considerazione anche il livello di personale conoscenza del digitale [3]. Infatti, venendo meno la stabilità del cartaceo, nel digitale le accortezze che un autore deve seguire per preservare il suo archivio si amplificano enormemente rispetto all'analogico. Ad un livello più implicito, possiamo quindi rintracciare la volontà d'archivio anche nell'attuazione di buone pratiche di gestione documentale, tali non tanto da garantire la conservazione a lungo termine, ma quantomeno limitare la possibilità di perdita dei propri dati.

La gestione documentale da parte singoli soggetti produttori, altrimenti detta *Personal Information Management* (PIM), è stata oggetto di analisi e sondaggi sin dai primi anni duemila [17] [19]. Se da un lato si è tentati dall'ipotizzare che gli scrittori, che talvolta vivono delle loro opere, prestino più attenzione ai loro prodotti digitali rispetto ai comuni cittadini, i risultati della ricerca mostrano come anche buona parte di loro abbia una pessima gestione dei loro archivi [3][11]. Questo atteggiamento, estremamente diffuso, è stato definito da Marshall, Bly e Brun-Cottan “benign neglect” [19], ossia un approccio fatalista e disimpegnato nei confronti del proprio patrimonio digitale. Si tratta di una tendenza solo in parte giustificabile con la mancanza di istruzioni e comprensione delle buone pratiche, dal momento in cui pare che la maggior parte degli scrittori non si preoccupi di ricercare in prima persona assistenza o informazioni per una corretta gestione documentale [3].

Ma se la gestione di file e cartelle può in qualche modo ricalcare la gestione di fascicoli cartacei, così come le e-mail sono equiparabili alla corrispondenza, il contesto digitale ha enormemente ampliato le forme di produzione documentaria, per cui ad oggi non è possibile parlare di *Personal Digital Archiving* senza menzionare la produzione di contenuti digitali su

social network, siti web e piattaforme di vario genere. Gli atteggiamenti degli utenti in relazione ai contenuti creati sui social media e siti web sono stati analizzati da molteplici prospettive come, ad esempio, privacy, comunicazione e impatto sociale [7] [14] [24]. Dal punto di vista dell'archiviazione, una recente indagine di Cannelli e Musso [4] ha evidenziato come gli utenti conferiscano un valore ai loro contenuti social, considerandoli parte integrante dei loro archivi personali. Tuttavia, risulta evidente come questo atteggiamento non sembri essere supportato da adeguate strategie di conservazione. Le autrici citano la presenza di un "implicit trust" che gli utenti ripongono nelle piattaforme di social media per quanto riguarda la conservazione a lungo dei propri dati, che li fa desistere dalla maturazione di concrete strategie di conservazione [4]. Ciò si scontra con l'assenza di veri e propri piani di conservazione delle piattaforme social, che anche quando contemplano il download dei propri contenuti, forniscono strumenti creati più con l'obiettivo della trasparenza sui dati, piuttosto che fornire un servizio di stampa archivistico [4]. Da questo punto di vista, il concetto di "implicit trust" nei social può essere letto come una nuova sfaccettatura della tendenza al "benign neglect" teorizzata nel 2006 da Marshall, Bly e Brun-Cottan [19].

La quantità di tempo e conoscenze necessarie per curare e gestire il proprio archivio – altamente frammentato e diffuso, online e offline – viene generalmente percepita come una sfida dai singoli soggetti [25]. La volontà d'archivio nel digitale risulta quindi ostacolata, in un momento storico in cui, ben più di prima, sarebbe necessaria per la conservazione delle "carte". In linea generale, la ricerca ha dimostrato che gli utenti riconoscono l'inadeguatezza delle loro pratiche di gestione delle informazioni personali o di archiviazione digitale [2][3][11][20], ma se da un lato una buona parte sembra essere aperta al ricevere istruzioni, dall'altro i ricercatori riferiscono che molti individui acquistano più spazio di archiviazione digitale per mantenere archivi più grandi e non gestiti, piuttosto che dedicare tempo alla cura dei propri file [8][20]. A questo proposito, occorre tenere in considerazione che il punto di vista sugli archivi digitali degli autori da un lato e degli istituti culturali dall'altro non può convergere sulle stesse pratiche, poiché nella gestione documentale le motivazioni e le logiche che muovono gli uni e gli altri sono oggettivamente diverse [8], così come è sempre stato per secoli anche per la produzione cartacea. I processi creativi con cui vengono prodotti i file potrebbero semplicemente non essere compatibili con le procedure strutturate raccomandate per l'archiviazione digitale [8]. Il livello di attenzione nei confronti delle corrette pratiche digitali sembrerebbe quindi essere la risultante del rapporto – non sempre lineare – fra il flusso creativo e il grado di competenza digitale di un autore.

La gran parte degli studi conclude appellandosi alla necessità dell'intervento delle istituzioni nell'educazione al *Personal Digital Archiving* [3] [11] [27], che effettivamente negli anni si sono prodigate in numerosi progetti di formazione, con la Library of Congress in prima fila<sup>1</sup>. Le procedure raccomandate per l'archiviazione digitale proposte dalle istituzioni sono, tuttavia, piuttosto complicate e, per quanto ideali dal punto di vista archivistico, sono difficilmente perseguibili nella quotidianità. In particolare, i progetti retrospettivi che richiedono all'autore di fare un censimento di tutti i suoi materiali digitali, per poi trattarli in linea con le best practices<sup>2</sup>, sono sicuramente adottabili dagli scrittori muniti di tempo e buona volontà, ma non sono realisticamente applicabili su larga scala. Utilizzando le parole di Marshall, infatti, possiamo chiederci «how relevant are these best practices to the consumer at home who has neither the resources, inclinations, skills, nor time to apply them?» [20].

### 3. VADEMECUM

Le istituzioni culturali non possono affidarsi totalmente alla gestione documentale dei singoli – della quale non potranno mai avere pieno controllo – per l'ottenimento di contenuti e infrastrutture digitali sostenibili dal punto di vista conservativo. Da questo punto di vista, la ricerca dovrebbe tendere i suoi sforzi verso lo sviluppo di tecnologie di acquisizione, gestione e conservazione flessibili, in grado di dialogare con contenuti, formati e strutture altamente variabili. Ciò non deresponsabilizza le istituzioni del loro ruolo educativo, ma la formazione deve essere ridimensionata verso buone pratiche "minimal", che non comportino un'eccessiva curatela autoriale e non interferiscano con il processo creativo. Alla luce di questa considerazione e delle precedenti iniziative di formazione, potremmo individuare un nucleo informativo di base sintetizzabile nei seguenti punti:

1. **Formati.** Si consiglia di utilizzare formati aperti, standard, non proprietari e indipendenti dal dispositivo. Le estensioni dei formati consigliati sono:

- Per documenti di testo: .docx, .dotx, .odt, .pdf, .pdf/a

<sup>1</sup> <https://digitalpreservation.gov/personalarchiving/>

<sup>2</sup> Ad esempio, le prime due fasi del "five step model" riportato da diverse istituzioni, come la University of Michigan Library. Cfr. <https://digitalpreservation.gov/personalarchiving/>

- Per fogli di calcolo: .xlsx, .xltx, .ods, .odp
- Per presentazioni: .pptx, .ppsx, .potx, .pdf, .pdf/a
- Per immagini (raster): .png, .tiff, .tif, .jpg, .exr, .mxf, .amf, .clf
- Per immagini (vettoriali): .svg, .dwg, .dwt
- Per audio e musica: .wav, .bmf, .rf64, .flac, .pcm, .raw, .sam, .musicxml
- Per video: .mpeg, .mpg, .mp2v, .vc3, .wav
- Per e-mail: .eml, .mbox,
- Per ipertesti: .xml, .html, .xsd, .xsl, .xslt, .css, .md
- Per database: .sql, .json, .jsonld, .csv
- 

Se l'estensione di un file prodotto o acquisito non rientra nelle casistiche sopraelencate, si consiglia di modificarla di conseguenza, conservando comunque il file originale.

2. **Denominazione.** Per la denominazione di file e cartelle, si consiglia di:

- Utilizzare parole chiave uniche, coincise e descrittive del contenuto del file.
- Utilizzare denominazioni brevi (possibilmente minori di 25 caratteri).
- Evitare di utilizzare caratteri speciali quali “.”, “,”, “%”, “\$”, “@”, “;”.
- Evitare di utilizzare spazi fra le parole chiave, prediligendo l'uso di underscore (ad es. “prova\_denominazione”) o la combinazione di lettere maiuscole e minuscole (ad es. “ProvaDenominazione”, “provaDenominazione”).
- Se si producono più files relativi allo sviluppo di uno stesso lavoro, si suggerisce di adottare un criterio univoco per segnalare la versione del file (ad es., “data\_nomefile”, “nomefile\_numero”, “nomefile\_rev\_numero”).

3. **Struttura.** Per impostare la struttura di cartelle del proprio archivio, si consiglia di:

- Creare una directory (cartella) principale, entro cui posizionare il resto delle cartelle. La directory principale, per immediatezza, può coincidere con il proprio desktop.
- Creare sottocartelle della directory principale seguendo il criterio di organizzazione dei file che si ritiene migliore per il loro utilizzo e recupero (ad es. per tema, attività, evento o tipologia documentaria)
- Evitare di creare strutture di cartelle a più di tre o quattro livelli.
- Evitare la creazione di file duplicati, anche se in cartelle o supporti diversi.

4. **Web archiving**

**Socialnetworks.** Qualora sia di interesse salvaguardare la propria produzione di contenuti social, diversi social network permettono di scaricare una copia dei dati presenti sulle rispettive piattaforme. Per farlo, si consiglia di:

- Individuare l'area dedicata al download dei propri dati sulla piattaforma social, generalmente situata nelle impostazioni generali o nelle impostazioni di privacy del profilo.
- Effettuare il download dei propri dati e archiviare il relativo file in un'area dedicata all'interno della struttura di cartelle della propria directory.

**Pagine web.** Qualora sia di interesse salvaguardare una o più pagine web di un sito, è possibile:

- Utilizzare servizi online gratuiti come Wayback Machine<sup>3</sup>, curato da Internet Archive e ampiamente utilizzato anche dalle istituzioni.
- Salvare la pagina web direttamente dal proprio browser e archiviare il relativo file in un'area dedicata all'interno della struttura di cartelle della propria directory.

5. **Conservazione e back-up.**

**Supporti.** A seconda delle esigenze di conservazione, è possibile fare riferimento ad uno o più supporti:

---

<sup>3</sup> <https://archive.org/web/>

- **Hard Disk (HD).** Effettuare backup su HD è una delle scelte più raccomandate per la sicurezza dei propri dati. Nella scelta della tipologia dell'HD, è consigliabile orientarsi verso le cosiddette SSD (Solid state drive): sono veloci, durature, leggere e maneggevoli. Un'alternativa alla SSD è l'hard drive disc (HDD): più economico di una SSD, è generalmente affidabile ma possiede parti mobili che potrebbero usurarsi in caso di ripetuto utilizzo. Per preservare al meglio un HD, è consigliabile evitare di sottoporlo a temperature estreme, cambiamenti di umidità, luce diretta, polvere o sporcizia.
- **Servizi Cloud.** In alternativa o in parallelo all'utilizzo di un HD, l'archiviazione online presenta diversi vantaggi a seconda dei provider: il servizio è gratuito entro un certo limite di Gb; spesso è possibile lavorare direttamente sul cloud e avere la versione più aggiornata di file di lavoro in tutti i dispositivi abilitati all'accesso; l'accesso ai file archiviati può avvenire in qualunque momento e luogo dotato di connessione internet. Tuttavia, presenta anche degli svantaggi da due principali punti di vista: è un servizio privato e, in quanto tale, è suscettibile di scelte di terzi in termini di continuità del servizio; è un servizio online, dunque soggetto a potenziali interruzioni, problemi di natura tecnica o attacchi hacker.
- **Chiavetta USB.** Limitare l'utilizzo di chiavette USB ai soli fini di trasferimento di files da un supporto di lavoro ad un altro, non come dispositivi di back-up. Per preservare al meglio una USB, è consigliabile evitare di sottoporla a temperature estreme, cambiamenti di umidità, luce diretta, polvere o sporcizia.
- **PC.** Se si conservano file rilevanti anche su smartphone o tablet, è consigliabile effettuare un back-up dei dispositivi sul proprio computer di riferimento. A questo punto, a loro volta, potranno essere integrati nelle pratiche generali di back-up.
- **E-mail.** Nonostante la pratica di auto-inviarsi files via mail sia molto diffusa, non è la tecnica migliore né per conservare i propri files né per averli prêt-à-porter. Nel primo caso, si consiglia di rivolgersi ad una delle soluzioni sopradescritte; nel secondo, si consiglia di utilizzare servizi Cloud, per poter avere a portata di mano su più dispositivi la stessa versione di un file.

## Mantenimento

- Si consiglia di effettuare il backup dei propri dati su due supporti da conservarsi, preferibilmente, in sedi separate.
  - Si consiglia di effettuare un backup dei propri dispositivi ad intervalli il più possibile regolari, a seconda delle proprie esigenze.
  - Qualsiasi sia la tipologia di storage prescelta, si consiglia di verificare annualmente il suo corretto funzionamento.
  - Si consiglia di aggiornare la tipologia di storage ogni 5-7 anni per prevenire la perdita di dati a causa dell'obsolescenza tecnologica.
6. **Successione.** Gli archivi digitali di persona sono, con ogni probabilità, la fonte più fragile del ventunesimo secolo. Tuttavia, se d'interesse, esistono diverse opzioni per consegnare il proprio archivio digitale ai posteri, siano essi familiari o una comunità di studiosi.
- **Istituzioni culturali.** Se si desidera donare una copia del proprio archivio per la ricerca, è possibile contattare un ente culturale e manifestare il proprio interesse ad effettuare un conferimento di materiali digitali. Qualora l'istituzione prescelta non sia disponibile, saprà consigliare enti più adatti o tecnicamente più qualificati per accogliere materiali digitali.
  - **Disposizioni testamentarie.** Per semplificare la gestione dell'eredità digitale, la soluzione migliore è quella di redigere un vero e proprio testamento contenente l'elenco analitico di tutti i beni componenti il proprio patrimonio digitale, inclusi gli account. Per consentire l'accesso ai legittimi eredi del materiale che dovesse trovarsi protetto da password, è necessario individuare un "legato di password" o redigere un "mandato post mortem" presso il proprio notaio di fiducia.

## 4. CONCLUSIONE

In un'epoca in cui la "volontà d'archivio" sembra essersi attenuata, la responsabilità di allontanarsi dalla paventata "Digital Dark Age" ricade principalmente sulle spalle delle istituzioni. Nel 2020 la *Open Preservation Foundation*<sup>4</sup> ha elaborato un sondaggio per analizzare lo stato dell'arte della *digital preservation* nelle istituzioni culturali [21] da cui è emersa una generalizzata necessità di supporto nell'elaborazione di *policies* e strumenti. In Italia – a livello normativo – sono stati fatti grandi passi avanti per la gestione della documentazione amministrativa, con l'introduzione del Codice dell'Amministrazione Digitale (CAD)<sup>5</sup> e delle Linee Guida AgID<sup>6</sup>, ma la necessità di strutturare la salvaguardia di altre tipologie di beni *born-digital* sembra essere passata in sordina. Per preservare il digitale d'autore occorre allora procedere ad una sensibilizzazione su più fronti:

- Informando gli autori sulle buone pratiche minime di gestione documentale, attraverso processi di apprendimento "smart" che non interferiscano eccessivamente con il processo creativo.
- Sensibilizzando archivisti e operatori delle istituzioni culturali che dovranno sempre più spesso affrontare acquisizioni di archivi ibridi o digitali.
- Lanciando campagne e progetti di acquisizione, volti all'intercettazione di archivi digitali potenzialmente interessanti per la ricerca e al loro conferimento.
- Investendo energie e fondi sullo sviluppo di standard e software condivisi per la gestione istituzionale dell'intero workflow di documenti nativi digitali d'autore.

A valle del processo di acquisizione, infatti, si aprono ulteriori interrogativi cruciali, quali il rispetto delle leggi sulla riservatezza, sul diritto d'autore e, dunque, sulla consultabilità dei documenti, così come sul modello descrittivo più adatto a rappresentare la complessità di questo fenomeno documentario. Si tratta di quesiti che, a loro volta, possono risultare ostativi all'accesso alla documentazione, ma non possono essere fattori bloccanti per l'acquisizione e la ricerca. Al contrario, devono rappresentare uno stimolo per l'adozione e lo sviluppo di nuovi strumenti e modelli, con uno sguardo particolare verso l'adozione di un'AI etica, trasparente ed integrata alla metodologia archivistica.

## BIBLIOGRAFIA

- [1] Albonico, Simone, e Niccolò Scaffai, a c. di. *L'autore e il suo archivio. Archivi letterari lombardi del Novecento*. ArchiLett900, II. Milano: Officina libraria, 2015.
- [2] Alon, Lilach, e Rafi Nachmias. «Gaps between actual and ideal personal information management behavior». *Computers in Human Behavior* 107 (2020): 106292. <https://doi.org/10.1016/j.chb.2020.106292>
- [3] Becker, Devin, e Collier Nogue. «Saving-Over, Over-Saving, and the Future Mess of Writers' Digital Archives: A Survey Report on the Personal Digital Archiving Practices of Emerging Writers». *The American Archivist* 75, fasc. 2 (15 novembre 2012): 482–513. <https://doi.org/10.17723/aarc.75.2.t024180533382067>
- [4] Cannelli, Beatrice, e Marta Musso «Social media as part of personal digital archives: exploring users' practices and service providers' policies regarding the preservation of digital memories». *Arch Sci* 22, (2022): 259–283. <https://doi.org/10.1007/s10502-021-09379-8>
- [5] Carbé, Emmanuela. *Digitale d'autore. Macchine, archivi, letterature*. Siena: Firenze University Press - USiena Press, 2023. <https://doi.org/10.36253/979-12-215-0023-3>
- [6] Carrière, Jean-Claude. *Non sperate di liberarvi dei libri*. Milano: La nave di Teseo, I Delfini, 2017.
- [7] Chen, Zhen Troy e Ming Cheung. «Privacy perception and protection on Chinese social media: a case study of WeChat». *Ethics Inf Technol*, (2018): 279–289. <https://doi.org/10.1007/s10676-018-9480-6>
- [8] Condron, Melody. «Identifying Individual and Institutional Motivations in Personal Digital Archiving». *Preservation, Digital Technology & Culture (PDT&C)* 48, no. 1 (2019): 28-37. <https://doi.org/10.1515/pdte-2018-0032>
- [9] Cosenza, Giovanna. «Eco: dal Medioevo al mondo digitale». *Il Libro dell'Anno* (2016). [https://www.treccani.it/enciclopedia/ecco-dal-medioevo-al-mondo-digitale\\_\(Il-Libro-dell'Anno\)](https://www.treccani.it/enciclopedia/ecco-dal-medioevo-al-mondo-digitale_(Il-Libro-dell'Anno))
- [10] Del Vento, Christian, e Pierre Musitelli. «La tradition italienne des manuscrits d'auteur : un patrimoine préservé et une culture littéraire». *Genesis. Manuscrits – Recherche – Invention*, fasc. 49 (15 dicembre 2019): 7–12.
- [11] Drosopoulou, Loukia, e Andrew Cox. «Information School Academics and the Value of Their Personal Digital Archives». *Information Research* 25, fasc. 3 (30 settembre 2020). <http://informationr.net/ir/25-3/paper872.html>

---

<sup>4</sup> <https://openpreservation.org/>

<sup>5</sup> <https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-03-07:82>

<sup>6</sup> <https://www.agid.gov.it/linee-guida>

- [12] Fiormonte, Domenico. «Antologia (e archeologia) della scrittura elettronica: tre tappe di un processo in corso», in C. Leonardi, M. Morelli, F. Santi, a c. di, *Modi di scrivere. Tecnologie e pratiche della scrittura dal manoscritto al CD-ROM. Atti del convegno di studio della Fondazione Ezio Franceschini e della Fondazione IBM Italia. Certosa del Galluzzo (Firenze), 11-12 ottobre 1996*. Spoleto: Centro di Studi Sull'Alto Medioevo (1997): 153-192.
- [13] Fiormonte, Domenico. *Scrittura e filologia nell'era digitale*. Torino: Bollati Boringhieri - Nuova didattica, 2003.
- [14] Hatfield, Elaine e Richard L. Rapson. «From pen pals to chat rooms: the impact of social media on Middle Eastern Society». *SpringerPlus* 4, 254, (2015). <https://doi.org/10.1186/s40064-015-1033-4>
- [15] Italia, Paola, a c. di. *A carte scoperte: come lavorano le scrittrici e gli scrittori contemporanei*. Bologna: Bononia University Press, 2021.
- [16] Italia, Paola, e Monica Zanardo. «Il Progetto – volontadarchivio», 2022. <https://volontadarchivio.disll.unipd.it/progetto/>
- [17] Jones, William e Jaime Teevan. «Introduction.» In *Personal Information Management*. Seattle: University of Washington Press, 2007.
- [18] Marshall, Brianna H., a c. di. *The Complete Guide to Personal Digital Archiving*. UK edition. London: Facet Publishing, 2018.
- [19] Marshall, Catherine C., Sara Bly e Françoise Brun-Cottan. «The Long Term Fate of our Digital Belongings: Toward a Service Model for Personal Archives». *Archiving Conference, Society for Imaging Science and Technology*, 2006, v. 1 (2006): 25–30.
- [20] Marshall, Catherine C. «Challenges and Opportunities for Personal Digital Archiving». I, *Digital*, Chicago: Society of American Archivists (2011): 90–114.
- [21] Open Preservation Foundation. *Findings Report of the 2019-2020 Digital Preservation Community Survey*. 2020. [https://openpreservation.org/wp-content/uploads/public/resources/surveys/2020/OPFCommunitySurvey\\_Report\\_v04\\_200907.pdf](https://openpreservation.org/wp-content/uploads/public/resources/surveys/2020/OPFCommunitySurvey_Report_v04_200907.pdf)
- [22] Purcell, Aaron D. *The Digital Archives Handbook: A Guide to Creation, Management, and Preservation*. Lanham: Rowman & Littlefield, 2019.
- [23] Schneider, Josh, Chance Adams, Sally DeBauche, Reid Echols, Callum McKean, Jessica Moran, e Dorothy Waugh. «Appraising, Processing, and Providing Access to Email in Contemporary Literary Archives». *Archives & Manuscripts* 47, fasc. 3 (11 agosto 2019): 305–26. <https://doi.org/10.1080/01576895.2019.1622138>
- [24] Seyfi, Murat, and Ayda Uzunçarşılı Soydaş. «The Relationship Between Autobiographical Memory and Social Media: Sharing Childhood Photographs on Social Media». *Global Media Journal (İstanbul)* 8.15 (2017): 57.
- [25] Sinn, Donghee, Sujin Kim e Sue Yeon Syn. «Personal Digital Archiving: Influencing Factors and Challenges to Practices». *Library Hi Tech* 35.2 (2017): 222–39. <https://doi.org/10.1108/LHT-09-2016-0103>
- [26] Vlassenroot, Eveline, Sally Chambers, Sven Lieber, Alejandra Michel, Friedel Geeraert, Jessica Pranger, Julie Birkholz, e Peter Mechant. «Web-Archiving and Social Media: An Exploratory Analysis». *International Journal of Digital Humanities* 2, fasc. 1 (1 novembre 2021): 107–28. <https://doi.org/10.1007/s42803-021-00036-1>
- [27] Zhao, Yue, Xiaomei Wu, X e Shanshan Li. «Perceived values to personal digital archives and their relationship to archiving behaviours: An exploratory research based on grounded theory». *Journal of Librarianship and Information Science*, (2023). <https://doi.org.ezproxy.unibo.it/10.1177/09610006231161327>



# L'applicazione del digitale alla valorizzazione della lingua come patrimonio culturale immateriale

Mariangela Giglio<sup>1</sup>

<sup>1</sup> Università di Siena, Italia – mariangela.giglio0@gmail.com

## ABSTRACT

La questione della conservazione del patrimonio culturale intangibile è diventata, a partire dagli anni '90, un tema centrale per le istituzioni culturali. Una volta superata la rigida dualità tra patrimonio tangibile e intangibile si è giunti alla consapevolezza della necessità di operazioni concrete per la salvaguardia del patrimonio linguistico come espressione chiave del patrimonio culturale immateriale. In questa direzione le tecnologie del digitale offrono una serie di strumenti innovativi per la conservazione del patrimonio linguistico e, non secondariamente, per una sua fruizione da parte di una comunità che si spera sempre più ampia. A tal fine si è partiti dalla definizione della specificità del materiale linguistico come parte dell'Intangible Cultural Heritage, arrivando a trattare le sfide, le opportunità e le soluzioni offerte dal digitale per il trattamento del materiale linguistico.

## PAROLE CHIAVE

Valorizzazione dell'Intangible Cultural Heritage; lingua come Intangible Cultural Heritage; digitale per i musei; lingua e digitale; musei sulla lingua

## 1. TRA MATERIALE E IMMATERIALE

Le istituzioni che, storicamente, si sono occupate di cultura hanno un compito apparentemente paradossale: rendere fruibili in forma tangibile beni in buona parte immateriali. La dicotomia tra tangibile e intangibile si manifesta particolarmente viva nell'ambito delle leggi per la salvaguardia del patrimonio culturale in cui si è assistito per lungo tempo a una maggiore attenzione per il patrimonio materiale a discapito di quello immateriale. Nell'ultimo trentennio<sup>1</sup>, tuttavia, si è assistito a un nuovo interesse per gli aspetti intangibili del patrimonio culturale, con la definizione di una terminologia e di piani di lavoro per la salvaguardia del patrimonio culturale intangibile (o, per usare la più nota espressione inglese, *l'Intangible Cultural Heritage*, spesso abbreviato in ICH). Si è lungamente discusso, anche in sedi istituzionali, se la lingua si possa considerare o meno parte di tale patrimonio immateriale e la questione della tutela delle lingue nell'ambito dell'ICH è stata percepita progressivamente come sempre più importante. Sebbene ad oggi la lingua non sia considerata come parte *dell'Intangible Cultural Heritage* in senso stretto<sup>2</sup>, essa viene riconosciuta come espressione di tale patrimonio e, coerentemente con un interesse sempre più marcato per il patrimonio culturale immateriale, si è assistito a un rinnovato interesse anche per il trattamento della lingua come parte dell'*Intangible Cultural Heritage*.

## 2. MUSEI DELLA LINGUA: LA DIFFICOLTÀ NELLA RESA

La peculiarità del tema lingua rispetto ad altri materiali museali costringe a porre l'accento sull'aspetto immateriale che il museo si trova a trattare, delineando così nuove caratteristiche fondative di questa tipologia museale. I musei della lingua sono tra le più recenti istituzioni del mondo museale ed è forse per questo che sono ancora tristemente poco riconosciuti nell'ambito della salvaguardia del patrimonio culturale immateriale. L'attenzione, però, sta progressivamente aumentando: oggi i musei della lingua rappresentano uno dei settori più importanti e in forte crescita dell'ambito degli ICH Museums. Tale operazione non è esente da difficoltà, al contrario, sono numerosi i fattori che impattano la riuscita dell'operazione di mettere "la lingua in mostra".

Esporre del materiale linguistico non è un'attività semplice per ragioni di ordine tanto teorico quanto pratico. In prima istanza la lingua è un tema divisivo: le lingue sono parte imprescindibile del patrimonio culturale intangibile e sono strettamente connesse al senso di identità di un gruppo o di una comunità. Permane, tra l'altro, un diffuso scetticismo da

---

<sup>1</sup> Va precisato che la questione dell'ICH viene sollevata ben prima che nell'ultimo trentennio: il problema della tutela del patrimonio culturale è implicitamente posto con la fondazione stessa dell'UNESCO nel 1948; e già dagli anni '70 si attivarono iniziative di studio sulle forme di tradizioni orali in diversi Paesi del mondo. Ciononostante, l'espressione "Intangible Cultural Heritage" viene usata per la prima volta solo nel 1982, nella conferenza UNESCO in Messico, con la prima formulazione di un documento di raccomandazioni nel 1989.

<sup>2</sup> La scelta dell'UNESCO di non includere la lingua come parte dell'ICH è stata messa in discussione, tra gli altri, da Smeets (2004) e Bernini (2014) che sottolineano il ruolo cruciale della lingua come espressione di patrimonio culturale immateriale.

parte di coloro che lavorano nell'ambito del patrimonio culturale immateriale nei confronti dei musei tradizionali, ai quali si imputa una eccessiva rigidità e un approccio smodatamente antiquato. In senso più ampio, inoltre, è la lingua in sé a essere un oggetto di difficile trattamento per la sua natura astratta e prevalentemente orale. Quest'ultimo aspetto ha storicamente spinto il museo a dedicare sezioni innaturalmente ampie delle sue esposizioni alla parte scritta della lingua, tralasciando, se non ignorando del tutto, gli aspetti dell'oralità e le lingue tramandate esclusivamente in via orale (Cannata et al., 2019). L'aspetto di intangibilità della lingua, che la rende a pieno diritto parte del patrimonio culturale immateriale di una nazione, è risultato non solo dell'immaterialità dell'oggetto, ma anche dal suo dinamismo, come acutamente sottolineato da Cristina Amescua (2013). La sfida dell'esposizione tangibile di materiale intangibile, già di per sé complessa, viene dunque ulteriormente complicata dalla mutevolezza della materia in oggetto, così che si corre il rischio di esporre un'immagine statica di un oggetto per sua stessa natura dinamico. Per via di questa stessa natura viva e dinamica dell'oggetto esposto, inoltre, le mostre sulla lingua richiedono continui aggiornamenti e integrazioni per essere attuali. Tale aggiunta di materiali, d'altro canto, crea spesso veri e propri "muri testuali" che impattano la leggibilità e scoraggiano l'utente, complicando ulteriormente il trattamento del materiale. L'effetto "wall of text" sembrerebbe inevitabile anche per l'impostazione tradizionale dei musei che tende a preferire il trattamento di materiali scritti e che peggiora ulteriormente quando si aggiunge la traduzione dei materiali esposti. C'è da considerare, infine, la sfida che comporta, per un museo, la scelta di esporre materiali facenti parte di patrimonio culturale immateriale: ogni museo può presentare solo documenti dotati di una loro fisicità ma, al contempo, deve tentare tramite quegli stessi materiali fisici di rappresentare patrimoni immateriali.

In questa difficile sfida di valorizzazione della lingua come bene immateriale si può, tuttavia, beneficiare delle potenzialità, altrettanto immateriali, offerte dalle nuove tecnologie.

### 3. DUE IMMATERIALITÀ A CONFRONTO: LINGUA E DIGITALE

Nonostante le critiche che, talvolta giustamente, investono l'ambito museale in virtù della sua eccessiva staticità, si può affermare che i musei non si siano sottratti alla sfida della rappresentazione digitale della lingua. Dell'ottantina di musei della lingua presenti nel mondo ben quindici sono digitali, senza contare l'elevato numero di siti web, portali e strumenti online che permettono, a vario titolo, di indagare ambiti relativi alla lingua. D'altronde i vantaggi derivanti dall'utilizzo di tecnologie moderne nel trattamento di beni culturali sono numerosi.

Tra gli aspetti positivi che, sicuramente, è necessario citare c'è una virtuale eliminazione di confini fisici, con l'allargamento di utenza che potenzialmente ne consegue, nonché la possibilità di lavorare in un regime di interoperabilità massimizzata, di comunicare con gli altri centri e permettere la creazione di percorsi interni di ricerca, indicizzazione e studio congiunti. La comunicazione delle istituzioni museali con i pubblici online, tra l'altro, parrebbe incrementare la visibilità dei musei stessi permettendo di attirare l'attenzione del pubblico verso alcune realtà meno conosciute.

L'implementazione di queste tecnologie (auspicabile per poter permettere un avvicinamento al mondo museale) può assumere diverse forme, che vanno dalla semplice creazione di un sito web, contenente le informazioni principali sul museo, fino allo sviluppo di ambienti di esposizione virtuali e immersivi. Non trascurabile l'utilizzo del digitale per una più agevole gestione museale come i processi di online ticketing, di conteggio elettronico dei visitatori e di tecnologie interne per l'interazione utente-opera, come i QR code, le audioguide o i touch screen. Stando alle ultime analisi disponibili dell'ISTAT e al Rapporto dell'Osservatorio Innovazione Digitale nei Beni e Attività Culturali del Politecnico di Milano, sebbene il processo di digitalizzazione del patrimonio culturale e dei servizi erogati stia continuando, la situazione generale presenta ancora ampi margini di miglioramento nel nostro Paese, pur considerando la brusca accelerata subito a seguito delle chiusure forzate a seguito della pandemia (ISTAT, 2022; Osservatorio Innovazione digitale del Politecnico di Milano, 2022).

Per quanto riguarda più specificamente il trattamento del tema lingua si è riscontrato come il digitale rappresenti spesso un mezzo valido per un'efficace valorizzazione. L'importanza dei nuovi media per lo storytelling non può più essere sottovalutata ed è oggetto di riflessione da parte di numerosi studiosi (Wyman et al., 2011; Alexander, 2011), tanto che ad oggi si parla specificamente di *digital storytelling*. D'altronde negli ultimi anni, l'attenzione degli studiosi è andata progressivamente aumentando relativamente l'evoluzione dei paradigmi narrativi crossmediali e transmediali (Jenkins, 2006), etichette che si affiancano e, talvolta impropriamente, si sostituiscono al concetto di *digital storytelling* (Meschini, 2018). L'importanza della narrazione è di per sé una componente centrale per l'esperienza museale e la tecnologia può porsi al servizio di questa esperienza sfruttando questa dimensione crossmediale e creando una narrativa che renda possibile fruire di oggetti immateriali tramite i mezzi, altrettanto intangibili, del digitale<sup>3</sup>. Uno degli strumenti che ha saputo di

---

<sup>3</sup> Particolarmente ben riuscita, in ambito italiano, l'esperienza del Museo archeologico di Napoli, che nel 2017 ha lanciato il Progetto "Father and son" un videogame che permette uno storytelling multimediale a partire dalle collezioni presenti nel museo. Il giocatore, nei

recente conquistare i favori delle istituzioni è sicuramente la tecnologia di realtà aumentata (*Augmented reality*, AR), che consente di sovrapporre diversi livelli informativi a luoghi concreti dando la possibilità, ad esempio, di creare più storie ambientate nello stesso luogo ma in momenti storici diversi, creando un senso di continuità e coinvolgimento per l'utente (Meschini, 2018). Altrettanto apprezzata sono le tecnologie di realtà virtuale (*Virtual Reality*, VR) che, tramite l'utilizzo di strumenti appositi, permettono l'immersione dell'utente in un ambiente virtuale. In ambito museale l'utilizzo di tecnologie di realtà aumentata e di realtà virtuale ha all'attivo già numerose esperienze, tanto che esistono diversi servizi<sup>4</sup> che permettono di sfruttare i dispositivi mobili degli utenti per dare vita agli oggetti esposti nel museo e creare un'interazione avvincente tra utente e opere esposte. La tecnologia AR, tuttavia, risulta particolarmente adeguata alle piccole realtà museali: più economica della (altrettanto evocativa) realtà virtuale, può essere una scelta vincente per convenire informazioni senza appesantire troppo le etichette dei materiali esposti. La narrazione museale, unitamente a queste tecnologie, può unire facilmente media diversi e consentire un'esperienza di fruizione completa e al contempo coinvolgente e piacevole per l'utente.

Particolarmente efficace per l'utenza dei musei della lingua è la possibilità di intrecciare storie collettive e storie personali. Questa tecnica di narrazione è fortemente coinvolgente per il pubblico e sembrerebbe particolarmente adatta a essere utilizzata per trattare il tema lingua, considerata la natura composita e sfaccettata del materiale linguistico.

Per la costruzione di un museo della lingua adeguato bisogna dunque unire le testimonianze documentali (comunque inevitabili) a percorsi più narrativi, utilizzando al meglio le possibilità offerte dal digitale. Il mondo museale (ma in senso più ampio quello delle istituzioni culturali) è d'altronde al centro di una trasformazione radicale dovuta all'esplosione del digitale. L'esito di questo cambiamento, almeno in potenza, dovrebbe essere quello di rendere il museo un luogo realmente inclusivo per ogni tipo di pubblico, implementando le nuove tecnologie per trasmettere un patrimonio culturale sempre più sfaccettato e multimediale.

#### 4. ESEMPI VIRTUOSI

Alla luce delle considerazioni sulla specificità del materiale linguistico come parte dell'ICH e delle sfide che esso presenta in termini di conservazione e valorizzazione, è opportuno esaminare alcune delle soluzioni offerte dal digitale per affrontare tali problematiche. In ambito museale un esempio virtuoso, non solo per l'utilizzo di tecniche di esposizione avanguardistiche ma anche per l'estensione e vitalità dei materiali esposti, è sicuramente quello del Museu da língua portuguesa. L'utilizzo dei nuovi media e tecnologie rappresenta uno dei punti forti del museo che ne rivendica esplicitamente l'utilizzo proprio in virtù del tema; per il museo, infatti, l'uso della tecnologia per l'elaborazione delle mostre è connesso al tema proposto: il patrimonio culturale immateriale<sup>5</sup>. Particolarmente interessante in tal senso è il Beco das Palavras: questa esposizione consiste in un gioco di etimologia interattiva in cui il visitatore può tentare, su schermi dotati di sensori di movimento, di unire pezzi di parole per crearne una intera e imparare la sua etimologia. Ancora la tecnologia è parte imprescindibile della mostra/gioco Palavras Cruzadas, che parte dall'utilizzo di totem interattivi per offrire la possibilità di "giocare" un doppio percorso di scoperta dell'apporto di altre lingue al portoghese. Con il percorso "parole" si può risalire all'origine del prestito e alla pronuncia nella lingua di partenza mentre il percorso "persona" permette di scoprire la cultura e le tradizioni dei migranti arrivati in Brasile, con la ricchezza del loro bagaglio linguistico. L'impiego della tecnologia, come visto, ben si sposa con la scelta di un approccio ludico alla conoscenza, che permette di mostrare la lingua in un'ottica di viaggio didattico sensoriale e soggettivo. Particolarmente riuscito, ad esempio, è il tentativo di mostrare la lingua come elemento al contempo legato al singolo individuo e alla collettività, come nell'allestimento digitale della "foresta sonora di lingue". L'esperienza consta in una serie di audio contenenti frasi tradotte in portoghese da varie lingue, rappresentate come rami di un'unica foresta. I rami-lingua sono ben settemila, corrispondenti grossomodo al numero di lingue del mondo. Il museo si presenta, quindi, come uno spazio sia tradizionale, ma con un ampio spazio dedicato all'interattività e alla multimedialità. D'altronde il ripensamento del museo è un'incombenza non ulteriormente procrastinabile e anche i musei più tradizionali, oltre quelli virtuali, sono obbligati a riconsiderare le modalità di comunicazione e divulgazione in base alle nuove necessità dell'utenza (Pescarin, 2014).

In linea con queste esigenze di ripensamento è l'approccio proposto dal Mundolingua, il grande museo parigino dedicato al linguaggio, che si propone di richiamare l'attenzione sui temi della linguistica e delle scienze della lingua mantenendo un approccio il più possibile ludico. Al termine del percorso espositivo più tradizionale, che si compone di una serie di sale

---

panni di un ragazzo alla ricerca del padre, si muove a ritroso nella storia dai giorni nostri fino all'antica Roma, esplorando così le varie collezioni del museo in modo interattivo.

<sup>4</sup>Applicazioni come Smartify, Artivive, e Hoverly, oltre che, ovviamente, Google Arts and Culture, permettono, tramite la tecnologia AR di aggiungere informazioni a elementi quali opere d'arte o manufatti museali, offrendo esperienze immersive agli utenti dei musei.

<sup>5</sup>Per ulteriori informazioni sulle mostre, in continuo aggiornamento, è possibile consultare il sito web del Museu da língua portuguesa, disponibile online a <https://www.museudalinguaportuguesa.org.br/> (ultima consultazione, aprile 2023).

sulla storia del linguaggio e delle lingue del mondo, sono infatti presenti due sale particolarmente originali: la prima, un grande spazio di gioco in cui si esprime al meglio la vocazione ludica del museo, offre allestimenti con lingue inventate, giochi di parole, proverbi, slang e crittografie di ogni tipo. La seconda sala, invece, è dedicata alle nuove tecnologie legate al linguaggio, come la traduzione automatica da parte di software e sistemi di riconoscimento vocale. Entrambe le sale sfruttano ampiamente l'utilizzo di materiali multimediali per permettere di massimizzare l'interazione utente e museo (Cossé, 2015).

In ambito italiano, sebbene il progetto non sia ancora concluso, si prefigura particolarmente interessante il progetto del MUNDI, il MUseo Nazionale dell'Italiano, che si prefigge di coinvolgere i suoi utenti tramite un approccio che unisca didattica, gioco e tecnologia. Coerentemente con la vocazione innovativa voluta dai suoi curatori alla parte espositiva più tradizionale corrisponderà, dunque, una parte multimediale più innovativa, in cui sono presenti dispositivi tecnologici di vario tipo come schermi interattivi e teche digitali. Alcune di queste tecnologie sono state mostrate nella mostra introduttiva del museo, con postazioni audiovisive dedicate a temi cardine della storia della lingua italiana e totem interattivi per permettere agli utenti di rispondere a una serie di quesiti su alcune parole dell'italiano. Anche se non ancora realizzato, sembra molto promettente il proposito di seguire, per il nascente museo, un'impostazione capillare. I curatori, infatti, si propongono di seguire il modello del Diffuseum,<sup>6</sup> ossia un museo diffuso, in più luoghi, che ricerca tracce linguistiche e le collega anche in posti diversi. Un modo per farlo potrebbe essere quello di usare la tecnologia AR e riunire così le tracce linguistiche ai luoghi e viceversa<sup>7</sup>, per creare una rete di informazioni. Questo approccio, interessante per ogni museo della lingua, risulta particolarmente adeguato nel caso specifico del MUNDI, in quanto la sua collocazione geografica (Firenze, all'interno del complesso di Santa Maria Novella) consentirebbe facilmente di collegare diversi poli informativi rilevanti per la storia linguistica italiana.

## 5. ALCUNE CONSIDERAZIONI

Le nuove potenzialità offerte dal digitale, sebbene preziose ai fini del trattamento dei beni culturali da parte delle istituzioni culturali, non sono purtroppo da sole garanzia di successo. L'utilizzo di questo strumento non è infatti esente da alcune criticità proprie di un mezzo non neutro. Alcune riflessioni rimangono, ad oggi, necessarie. In primis la questione tecnologica, strettamente correlata in alcune sue applicazioni a quella del divario economico. La diffusione delle tecnologie digitali, infatti, pur permettendo teoricamente un accesso a fasce più ampie di popolazione, può accompagnarsi a un aumento di tale dislivello. È il caso dell'approccio cosiddetto "BYOD" (*Bring your own device*) per cui all'utente è richiesto di utilizzare il proprio dispositivo (solitamente lo smartphone o il tablet) per interagire con i contenuti proposti. Permangono, ovviamente, i problemi relativi al cosiddetto "*digital gap*" e alla "*digital illiteracy*" che ancora affliggono la nostra società e rimangono inalterate, per di più, le difficoltà di accesso a tali risorse da parte di coloro i quali risiedono in territori in cui l'accesso a Internet è soggetto a limitazioni.

Nel pensare a un museo proiettato al digitale, anche se si parte con le migliori intenzioni di inclusione, si corre dunque il rischio di lasciare esclusa una ampia fetta di popolazione. Nel contesto italiano il pensiero va, ad esempio, agli utenti più anziani e più inesperti con la tecnologia ma anche, sorprendentemente, alle nuovissime generazioni che hanno spesso una conoscenza tecnologica imperfetta e che necessita di essere ancora formata. Il problema sociale, sempre rilevante, assume particolare importanza quando investe l'ambito linguistico, per le ovvie motivazioni connesse al valore della lingua quale patrimonio comune. Persistono, inoltre, tutte le difficoltà connesse alla resa dell'immateriale nel mondo digitale. L'argomento, sebbene non nuovo alla museologia contemporanea, rimane purtroppo ancora poco trattato nelle sue implicazioni pratiche.

Rimane infine da considerare il problema della capacità effettiva, da parte del museo, di convogliare efficacemente le informazioni. Questa operazione è vitale se si vuole creare uno spazio museale che sia veramente aperto a tutti e non solo a coloro che sono in grado di comprendere e interagire con la complessità delle opere esposte e dei percorsi proposti. Questo è dunque lo spazio di azione delle nuove tecnologie, che andranno sfruttate come strumenti per avvicinare e

---

<sup>6</sup> Diffuseum, o museo diffuso, è il nome della realtà museale elaborata nell'ambito del progetto "Eurotales: Museo interattivo delle lingue d'Europa" curato da Nadia Cannata. Principio chiave del diffuseum è il riconoscere la presenza di tracce linguistiche (quali monumenti, oggetti, graffiti, iscrizioni ecc.) come elementi, attuali o passati, testimoni di una lingua. Il Diffuseum lega tra loro queste tracce creando dei percorsi investigativi unici nel tempo e nello spazio, consentendo una visione di insieme delle tracce che superi i confini storico/geografici. Per ulteriori informazioni è possibile consultare il sito <http://eurotales.eu/en/home-english/> (ultima consultazione novembre 2022).

<sup>7</sup> Al di fuori dell'ambito strettamente linguistico un progetto simile è stato utilizzato dal sistema di navigazione AR usato a Pompei. Il Sistema AR per la ricostruzione digitale permette ai visitatori di camminare per il sito archeologico e di scoprire informazioni sul luogo utilizzando la fotocamera del proprio dispositivo. Tale tecnologia potrebbe risultare altrettanto promettente in ambito linguistico, legando, ad esempio, i luoghi di produzione di documenti linguistici alle informazioni contenute nel museo e, viceversa, il museo ai suoi documenti, così da unire luoghi di produzione e luoghi di conservazione.

coinvolgere un'utenza quanto più possibile ampia senza incorrere nella tentazione della tecnologia come mezzo fine a sé stesso.

## 6. CONCLUSIONI

Da quanto emerso finora appare evidente che, sebbene si stia andando nella giusta direzione, molta strada rimane ancora da fare nell'ambito di un trattamento efficace della lingua come bene culturale immateriale. Queste considerazioni imporrebbero di agire subito e avviare senza indugio iniziative atte a una seria riconsiderazione di alcune delle metodologie più antiquate di trattamento dei beni culturali, con un ripensamento di metodi espositivi e materiali.

Sfortunatamente su questo punto le istituzioni culturali si trovano ancora troppo spesso ancorate a un approccio eccessivamente rigido e incapace di valorizzare adeguatamente i materiali esposti. Tali problemi sono imputabili a una certa reticenza di principio ma anche, non banalmente, a difficoltà economiche e organizzative che impattano tanto più duramente le piccole realtà, impossibilitate a intraprendere percorsi di questa natura. A tal proposito si è notato inoltre come la scelta di lasciare che il costo delle implementazioni tecnologiche ricada sull'utente finale (come nel caso dell'utilizzo di tecnologie BYOD), sebbene teoricamente efficace, possa avere delle ripercussioni sulle fasce più indigenti della popolazione. La questione, che è complessa e sfaccettata, rimane ancora aperta. Non mancano, tuttavia, iniziative virtuose che hanno saputo, in diversa misura, implementare le più recenti tecnologie al servizio dell'utente finale, ed altre ancora in fieri che si spera possano riuscire nell'obiettivo.

## BIBLIOGRAFIA

- [1] Alexander, Bryan. *New Digital Storytelling: Creating Narratives with New Media*. ABC-CLIO, LLC, 2017.
- [2] Amescua, Cristina, e Arizpe, Lourdes. *Anthropological Perspectives on Intangible Cultural Heritage*. Springer, 2013.
- [3] Bernini, Andrea. «Languages as intangible cultural heritage: about an ecolinguistic capital». *Sustainable Multilingualism* 5 (2014). <https://doi.org/10.7220/2335-2027.5.6>
- [4] Cannata, Nadia, Maia Wellington, Gahtan e Margaret J. M. Sönmez. *Museums of Language and the Display of Intangible Cultural Heritage*. Taylor & Francis Group, 2019.
- [5] Cossé, Pierre-Yves. «Mundolinga, la découverte du monde des langues». *La Tribune*, 17 agosto 2015.
- [6] Grepstad, Ottar. *Language museums of the world*. Ottar Grepstad and Centre for Norwegian Language and Literature, 2018.
- [7] ISTAT, *Report Musei e Istituzioni Similari*, 2020
- [8] Jenkins, Henry. *Convergence culture: Where old and new media collide*. New York University Press, 2006
- [9] Mandarano, Nicolette. *Musei e media digitali*. Carocci, 2019.
- [10] Meschini, Federico. *Reti, memoria e narrazione: archivi e biblioteche digitali tra ricostruzione e racconto*. Sette città, 2018.
- [11] Osservatorio Innovazione digitale del Politecnico di Milano. *L'innovazione digitale nei musei italiani nel 2022*. Politecnico di Milano, 2022
- [12] Pescarin, Sofia. «Museums and Virtual Museums in Europe: Reaching expectations». *SCIRES-IT: SCientific RESearch and Information Technology* 4 (2014). <http://doi.org/10.2423/i22394303v4n4p131>
- [13] Smeets, Riëks. «Language as a Vehicle of the Intangible Cultural Heritage». *Museum International* 56, n. 1-2 (maggio 2004): 156–65. <https://doi.org/10.1111/j.1350-0775.2004.00470.x>
- [14] UNESCO, *Convention for the safeguarding of intangible cultural heritage*. 2003. <https://ich.unesco.org/en/convention>
- [15] Wyman, Bruce, Scott Smith, Daniel Meyers e Michael Godfrey. «Digital Storytelling in Museums: Observations and Best Practices». *The Museum Journal* 54, 4 (ottobre 2011): 461–68. <https://doi.org/10.1111/j.2151-6952.2011.00110.x>
- [16] Yoshida, Kenji. «The Museum and the Intangible Cultural Heritage». *Museum International* 56, n. 1-2 (maggio 2004): 108–112. <https://doi.org/10.1111/j.1350-0775.2004.00464.x>

# Digital Accrocchio: a computational image searching tool for social history

Tiago Luís Gil<sup>1</sup>

<sup>1</sup> University of Brasilia, Brazil – tiagoluigil@gmail.com

## ABSTRACT

Social history is fundamentally an attempt to include ordinary people in history. This inclusion, however, has always been challenging due to the little representativeness of these people in the sources, which has always provided a noteworthy "inductivist" or "empiricist" character to the works of this type of historical research. Nevertheless, how to tell the story of ordinary people without using countless of sources to seek the remaining pieces of evidence of a few subalterns? This research is inserted in this framework, in the search for studying the cotton spinners and weavers in colonial Brazil, between 1790 and 1810. Besides being done by subaltern people, this activity was partly forbidden due to mercantilist policies. Despite this, several records of its broad and significant existence exist in a documental set of about 12,000 handwritten documents available in Archives. To carry out this research, we will use technological resources of programming with python and openCV, in an attempt to create a tool that can scan thousands of digitised documents in the search for those (so far) unknown women.

## KEYWORDS

Computational research tools; social history; fabric production

## 1. INTRODUCTION

For centuries, history was the story of “great” men: kings, generals, and presidents. In the 1950s, social history arose to speak of peasants, serfs, and workers. This task implied the use of many sources because ordinary people have always been weakly represented in documents. In the 1950s, research began with punched cards that collected small fragments of information from each character to form a bigger picture.[1,2] In the 1980s, these researches started to use computers, and several methods and programs were created within this effort.[3,4,5] This proposal aims to start from a specific research problem and develop digital solutions tailored to those needs, in tune with those initiatives from the 1980s. The case study will be the captaincy of São Paulo (currently Brazil) between the end of the 18<sup>th</sup> century and the beginning of the 19<sup>th</sup>.

## 2. THE CONTEXT OF SPINNERS AND SOURCES FOR THEIR RESEARCH

At the end of the 18<sup>th</sup> century and the beginning of the 19<sup>th</sup> century, there was widespread spinning and weaving in several cities of the Captaincy of São Paulo, in Brazil. It was an activity done by women, primarily poor women and widows, among the humblest half of the population. The cloth production was made in combination by spinners and weavers, who occupied about 15% of the houses in the localities of the Paraíba valley. This activity was the second most practiced in those localities, surpassed only by agriculture.[6]

The production of cloth was prohibited in Brazil, due to a Decree of 1795 that, with the argument of avoiding the neglect of agriculture, aimed to reserve the colonial consumer market for portuguese middlemen of European fabrics in a context in which England was emerging as the world's workshop. Despite the general prohibition, there was room for weaving insofar as the production of thick fabrics used for making clothes for the enslaved people was permitted. There is no way to be sure (nor was there at the time) about the use of these fabrics by free people and in this interstice that the continuity of this production occurred.[7]

Weaving was not effectively combated, despite the prohibition. However, because of the Decree - which historians took exceptionally seriously - this theme was entirely ignored by historiography. At the same time, historiography guided the work of archivists, who never included terms such as weaving, spinning, and fabric production in the cataloguing sheets, hindering any research on this subject.

The documents available are the so-called nominative lists of inhabitants, a kind of census carried out for economic purposes that listed the occupation of each house. The registers describe the entire community house by house, indicating its residents, age, and marital status of each, as well as free and enslaved persons. Finally, each house has its main economic activity described, often in detail. The famous demographic historian Louis Henry addressed this particular document in the Brazilian editions of his famous "*Manuel de Dépouillement...*"[8]

The lists, however, form a very voluminous set of data that challenges the researcher. The classic works of social history ended up being restricted to a single city or community in such a way that it was possible to know all the anonymous people



in that locality. Such were the researches of Goubert and Ladurie, to give two classic examples, and it was also in community contexts that much microhistorical research was carried out. It was the necessary cutout, but also the cutout possible for the breath of a researcher, even with the help of many assistants, as was the case with Ladurie.[1,9, 10]

The research at hand intends to somewhat modify the scope of this approach. Instead of one community, four, Guaratinguetá, Pindamonhangaba, Taubaté and Lorena, all neighbouring and all with a strong emphasis on cloth production. The towns exchanged many people among themselves, as has been observed up to now, through marriages, cronyism, removals, and trade in enslaved people, much more than with other localities, indicating strong integration. Moreover, this integration was also fundamental so that the spinners could have the guarantee of a buyer of yarn since their poverty prevented any expectation. In other words, the inclusion of four communities is not an empirical documental increment but intends to show a geography of cloth production, which is only possible by taking the set of villages.

The question that arises is the volume of data that this four-village approach presupposes. If spinners appear in few documents, as such only in the lists, they appear as godmothers and mothers in baptismal records, wives, and witnesses in marriage registers, and buying and selling in notarial records. All this - which would already be much for a single locality - becomes an almost impossible amount of data for all localities together. Furthermore, it calls into question the viability of a single historian's research. It is not a matter, therefore, of speeding up or processing more documents but of making research possible, which starts with a question that demands little information spread over thousands of pages. For this, we intend to resort to the use of programming.

### 3. THE SOLUTION FOUND: A HOME-MADE TOOL

A document image management system of the villages studied in this research proposal was created. It is a system that organizes the images and is able, according to the specificity of each document, according to its visual morphology (whether in table form, as are the Nominative Lists, or in the form of text "nests", as are the baptisms) and separates them into records, house by house, in the case of the Lists, baptism by baptism, in the case of the parish registers. These documents are then inserted as a binary image into a sqlite database system (within a python installation) to be called up by a search system that identifies handwritten image patterns when required. This tool is not HTR software but a more straightforward solution that uses only a few text samples to search for specific words without the need for training via artificial intelligence – a mere word searcher or an image “ctrl + F”.



Figures. 1, 2 and 3. Source segmentation process. Figure 1. Original image. Figure 2. Original image with cut-off point. Figure 3. Cut image.

The available HTR programs are satisfying and worthy of their proposed tasks, but they would be impractical in such a research. First, segment recognition proved insufficient to account for the morphologies of documents made in the form of a register. It could not separate baptismal records from one another, nor the rows of a table. For this, the tools of OpenCV



(image management library available in many languages, including python) provided very efficient answers with a reasonably small and easy-to-manage code. The OpenCV library recognizes the table rows in the horizontal direction and establishes a cut-off point (Figure 2, middle column, with the pink line highlighted). This point is used to fragment the image, row by row or, in this case, house by house. The image recognition system scans each fragment in the search for specific terms.

HTR tools require a minimum amount of text (5000 words, for example), for which they promise an accuracy of over 90%. However, these high percentages are only possible with abundant "secondary" words (prepositions, articles, etc.). In contrast, many words with a high meaning impact in the text are often incorrectly transcribed because they were less common or due to machine learning overfitting.[11,12] In this case, a single example of the desired word (although a more significant number allows an increase in the accuracy of the search) is enough to find very close versions. A single image fragment (or "template", corresponding to a word), with a good calibration of OpenCV's matchTemplate tool, allowed finding 100% of similar occurrences (with a burden of 30% false positives) or 95% of similar occurrences without false positives. The use of another library, MTM MatchTemplates, which allows the simultaneous use of several "templates" (with variations of the same word), reached 100% of success with only 15% of false positives, having as a burden the need to insert three or four slightly different images of the same word (something effectively easy to accomplish). In all cases, a sample of 310 images was used. A traditional HTR tool would require training with thousands of words and would not effectively solve all problems.

Figures 4, 5, 6, 7, 8 and 9. Searching system and templates.

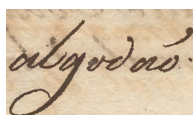


Figure 4. Search term: Cotton ("algodão").

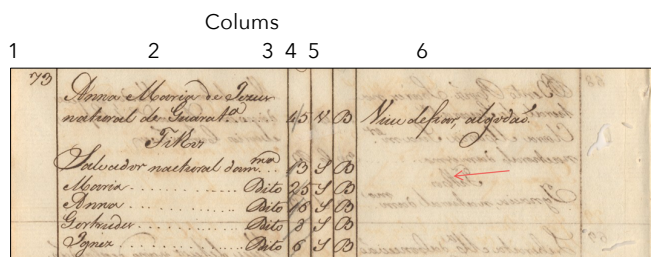


Figure 5. Original fragment taken from the image.

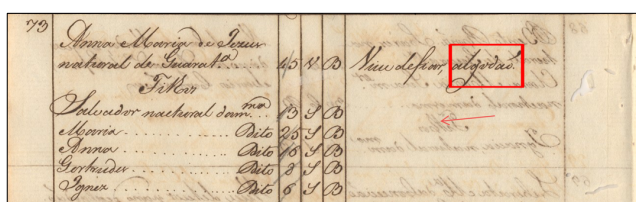


Figure 6 . Fragment with detected text.



Figure 7. Image processing illustration.

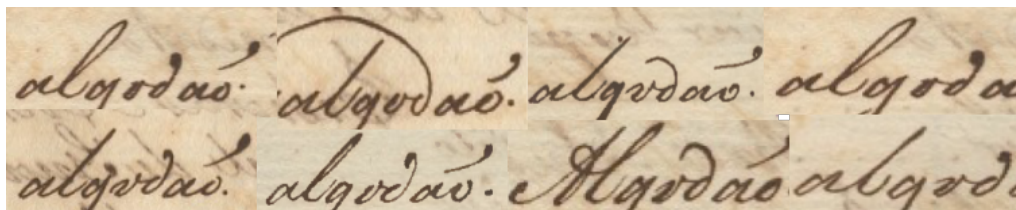


Figure 8. “Cotton” (“algodão”) variations in the same manuscript.

A source such as the nominative lists expresses this well: these are documents in the form of a table, with the first column being an identifying number and the second column being the names of the household members, followed by the column with their age, marital status (fourth column) and colour (fifth column). The sixth column contains information on the productive activities of the household. In the first column, there are only numbers; in the second column, only names follow specific patterns and have unique characteristics. It is not just a random text. Similarly, columns 3, 4 and 5 are filled only with letters and numbers, which is easy to inform the computer via python. These are unambiguous and simple rules. Finally, column 6 may count more word variations, but there is also a great regularity since many people did the same things, and the census takers repeat the same textual formula. That is a well-known regularity that can be easily transmitted to the computer with the aid of python and the OpenCV library. Furthermore, handwriting often varies in groups of 50 pages (as there are several census takers), and this would require new HTR training. The system developed by this research allows labour saving by dispensing with training and with good use of the materials produced.

#### 4. CONCLUSION

Different research has different demands on technology resources, and the craft of research can benefit from digital resources. However, large, robust solutions do not always bring direct benefits to all approaches, and often the needs of each research require ad-hoc answers. What is presented here is a specific solution for a very particular set of data, presented in the form of images of a document with its characteristics. Research focused on the history of ordinary people involves adequate treatment of the available material and the necessary data. Training with machine learning through some HTR model would not be possible given the amount of text available (often handwriting changes, and a minimum of 5000 words is often unattainable) and the characteristics of this text (names, ages and certain economic activities). In this sense, the use of the tools presented became quite opportune and allowed the continuity of the research without dismissing the digital resources that can help make it viable. Finally, it seems relevant to point out this example as a digital craft solution, with solutions tailored to the research and with great respect to the research problem presented, without imposing digital restrictions on the theory that guides the research.

#### REFERENCES

- [1] Ginzburg, Carlo. *Il formaggio e i vermi. Il cosmo di un mugnaio del '500*. Adelphi, 2019.
- [2] Goubert, Pierre. *Cent Mille Provinciaux au XVIIe Siècle. Beauvais et le Beauvaisis de 1600 à 1730*. Paris: Flammarion, 1968.
- [3] Thaller, Manfred. “Methods and techniques of historical computation”. Em *History and Computing*, edited by Peter Denley e Deian Hopkin. Manchester: Manchester University Press, 1987.
- [4] Genet, Jean-Philippe. “The PROSOP system”. Em *History and Computing*, edited by Peter Denley e Deian Hopkin. Manchester: Manchester University Press, 1987.
- [5] Harrison, Sarah, Charles Jardine, Jessica King, Tim King, e Alan Macfarlane. “Reconstructing Historical Communities by Computer”. *Current Anthropology* 20, no 04 (1979): 808–9.
- [6] *Listas Nominativas de Habitantes de Pindamonhangaba, Taubaté, Lorena e Guaratinguetá. 1800-1805*. Arquivo do Estado de São Paulo.
- [7] Novais, Fernando A. *Portugal e Brasil na crise do antigo sistema colonial (1777-1808)*. São Paulo: HUCITEC, 1989.
- [8] Henry, Louis. *Técnicas de análise em demografia histórica*. Curitiba: UFPR, 1977.
- [9] Ladurie, Emmanuel Le Roy. *Montaillou, village occitan de 1294 à 1324*. Paris: Folio, 2008.
- [10] Levi, Giovanni. *Centro e periferia di uno stato assoluto*. Torino: Rosenberg & Sellier, 1985.
- [11] Scheltjens, Werner. “The Feasibility of Machine-Learning Based Workflows for Editing Serial Historical Sources : First Results of the Schenkenschans Customs Registers Project”. Otto-Friedrich-Universität, 2023. <https://doi.org/10.20378/irb-58260>
- [12] Perdiki, Elpida. "Review of 'Transkribus: Reviewing HTR training on (Greek) manuscripts'." *RIDE* 15 (2022). DOI: 10.18716/ride.a.15.6. Accessed: 01.03.2023.

# Quali prospettive per ItaDraCor? Risorse e strumenti per la codifica di testi drammatici in lingua italiana

Luca Giovannini<sup>1</sup>, Ingo Börner<sup>2</sup>, Frank Fischer<sup>3</sup>, Carsten Milling<sup>4</sup>, Daniil Skorinkin<sup>5</sup>, Peer Trilcke<sup>2</sup>

<sup>1</sup> Universität Potsdam, Germania / Università di Padova, Italia – giovannini@uni-potsdam.de

<sup>2</sup> Universität Potsdam, Germania – börner@uni-potsdam.de

<sup>3</sup> Freie Universität Berlin, Germania – fr.fischer@fu-berlin.de

<sup>4</sup> Universität Potsdam, Germania – milling@uni-potsdam.de

<sup>5</sup> Universität Potsdam, Germania – skorinkin@uni-potsdam.de

<sup>5</sup> Universität Potsdam, Germania – trilcke@uni-potsdam.de

## ABSTRACT

Il contributo presenta brevemente DraCor, una delle maggiori infrastrutture a livello europeo per lo studio della letteratura teatrale, e gli strumenti computazionali attualmente in sviluppo per favorire l'inclusione di nuovi testi nei suoi corpora. Si focalizza poi sulla collezione in lingua italiana (ItaDraCor), di cui esamina il processo di creazione e l'attuale composizione, e ne tratteggia le prospettive future.

## PAROLE CHIAVE

Dramma; teatro; XML-TEI; corpora; DraCor

## 1. INTRODUZIONE

Attiva sin dal 2017, la piattaforma **DraCor** (<https://dracor.org>, TEI Rahtz Prize 2022) rappresenta una delle maggiori infrastrutture a livello europeo per lo studio della letteratura teatrale. Basata sul concetto di *programmable corpora* [2], che indica collezioni testuali accessibili ed interrogabili a più livelli, DraCor si colloca al crocevia tra filologia tradizionale e digitale, offrendo congiuntamente un ampio database di testi drammatici e una serie di risorse computazionali per la ricerca e la didattica. Allo stato attuale, il progetto dispone di 15 corpora in 13 lingue diverse, per un totale di circa 3000 opere, mentre ulteriori tre corpora sono a diversi stadi di produzione <sup>1</sup>. Ogni testo presente in DraCor è codificato in formato XML-TEI (P5), seguendo le linee guida specifiche per *performance texts* <sup>2</sup> e ponendo una particolare attenzione all'identificazione tramite tag dei personaggi parlanti. Pur mirando alla maggiore accuratezza possibile, DraCor non prepara edizioni critiche, ma aspira piuttosto a permettere l'indagine di ampi corpora testuali secondo i consolidati paradigmi del *distant reading* [5] e dello *scalable reading* [6].

DraCor offre un ampio ventaglio di possibilità per l'esplorazione delle sue collezioni. Un ruolo di prima importanza è svolto ad esempio dall'algoritmo per l'estrazione automatica di reti relazionali, che permette di investigare i *character networks* all'interno delle opere [3], ma è altresì possibile analizzare le relazioni tra i personaggi e la distribuzione delle battute [4]. L'API <sup>3</sup> permette inoltre di effettuare ricerche dettagliate nel contenuto dei drammi, estraendo ad esempio tutte le battute pronunciate da personaggi femminili in un'opera oppure le indicazioni di regia in un intero corpus [8]; l'uso di queste funzioni è facilitato da *wrapper* disponibili per R, Python e Haskell. La presenza di *Linked Open Data* nella codifica dei testi, infine, consente agili ricerche tramite SPARQL.

## 2. RISORSE E STRUMENTI

Fin dai suoi esordi, DraCor si propone come un progetto collaborativo e decentrato, sempre aperto a contributi di studiosi di diversa estrazione linguistica e culturale; non a caso, l'aggiunta di nuovi corpora è stata negli anni frutto sia di progetti interni sia di cooperazioni con istituzioni esterne. Nell'ottica di un consolidamento della comunità intorno a DraCor, e con l'intento di raggiungere nuovi potenziali contributori, sono attualmente in sviluppo diverse risorse e strumenti per rendere

---

<sup>1</sup> DraCor comprende corpora in alsaziano, baschi, francese, tedesco, greco antico, ungherese, italiano, latino, russo, spagnolo, svedese e tataro, oltre che tre collezioni dedicate all'opera di autori specifici (Calderón de La Barca, William Shakespeare e le traduzioni shakespeariane di August Wilhelm Schlegel e Ludwig Tieck). I corpora in produzione sono invece in lingua inglese, ucraina e ceca.

<sup>2</sup> <https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html>

<sup>3</sup> <https://dracor.org/doc/api>

il processo di *onboarding* ancora più accessibile – anche e soprattutto per ricercatori al di fuori dell’ambito delle *digital humanities*.

Il lavoro in questo senso procede su un binario parallelo: da un lato si sta provvedendo alla stesura di linee guida per formalizzare e omogeneizzare il processo di codifica, dall’altro sono allo studio metodi per semplificare la *pipeline* di trasformazione dei testi. Le collezioni drammatiche presenti in DraCor sono infatti di provenienza eterogenea, e le procedure variano in base alla qualità del loro markup originario: se basi di dati in altre varianti di XML possono essere adattate tramite script XSLT o Python ad hoc, opere con codifica limitata (es. in HTML) o assente (es. file TXT derivati da OCR) hanno posto finora maggiori sfide. Per facilitare questo compito è stato quindi sviluppato *ezdrama* [1], una semplice linguaggio di tipo *markdown* che può essere applicato tramite un qualsiasi editor testuale e permette (attraverso un programma Python <sup>4</sup>) la rapida conversione dell’opera in un file TEI-XML pronto per DraCor (Figura 1).

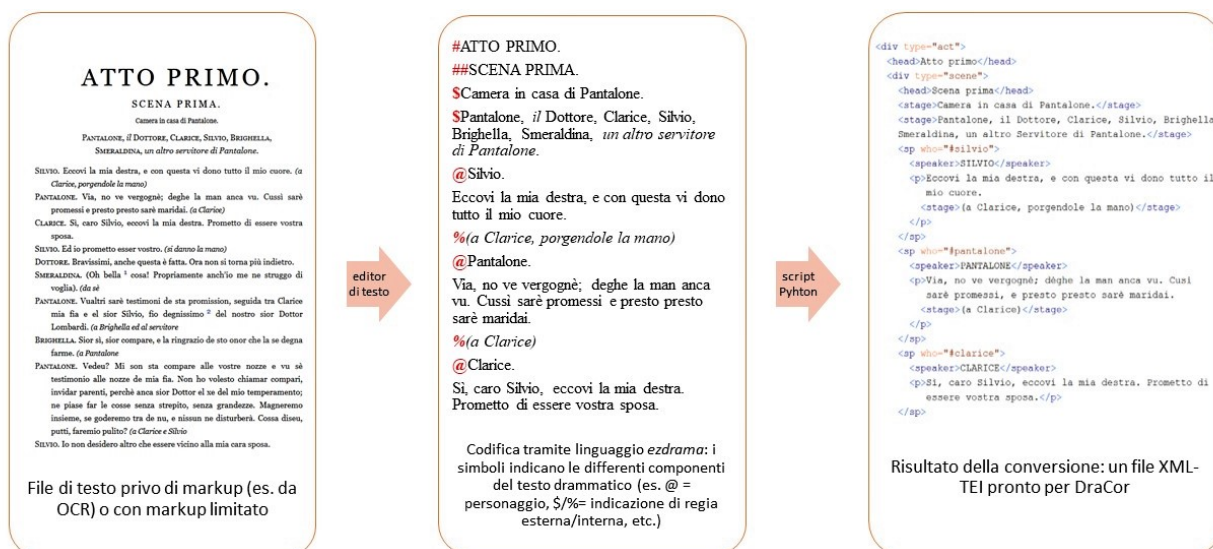


Figura 13. Workflow per l'utilizzo di *ezdrama*. L'esempio è tratto dall'incipit de Il servitore di due padroni di Carlo Goldoni (1745).

### 3. ITADRACOR: PANORAMICA E PROSPETTIVE

Il corpus drammatico in lingua italiana (*ItaDraCor*, <https://dracor.org/ita>) contiene al momento 139 opere originali, scritte o pubblicate tra il 1449 ed il 1933. Tutti i testi provengono dal progetto *Biblioteca italiana* [7], che offre libero accesso ad una vasta gamma di testi in edizione integrale e con codifica in XML/TEI (versione P4) <sup>5</sup>. Alla creazione di ItaDraCor (febbraio 2020), il catalogo della Biblioteca Italiana conteneva 171 elementi nella categoria "Letteratura teatrale", di cui 32 sono stati esclusi per ragioni pragmatiche (come la mancata identificazione dei personaggi) e/o letterarie (testi in latino o non propriamente drammatici). Il processo di conversione dei testi da Biblioteca Italiana a DraCor ha compreso anche una fase di arricchimento dei dati, con l’inserimento di ID Wikidata per autori, opere e personaggi storici o mitologici, di varie informazioni cronologiche (composizione, prima pubblicazione, prima rappresentazione) e di ulteriori metadati sui personaggi (come il genere), oltre che la correzione di errori derivati da precedenti OCR.

Dopo questa prima fase, tuttavia, lo sviluppo di ItaDraCor si è arrestato, e sebbene esso rappresenti ancora il quinto corpus per estensione, non è stato ampliato con nuovi testi né ha subito ulteriori miglioramenti nella codifica, fatta salva la manutenzione ordinaria. La letteratura teatrale del tardo Ottocento e del primo Novecento, in particolare, è pressoché assente, e opere di autori come Pirandello o Eduardo, oltre che quelle degli esponenti del teatro verista e delle Avanguardie storiche, attendono di essere aggiunte. Il lancio di *ezdrama* crea in questo senso nuove opportunità per l’ampliamento della collezione a partire da risorse digitali liberamente accessibili sul web ma prive di markup avanzato; un progetto pilota in tal senso riguarda, ad esempio, il vasto patrimonio di libretti d’opera disponibile in rete. <sup>6</sup> Nell’ottica di un rilancio di ItaDraCor, il team di DraCor intende quindi sfruttare l’occasione di AIUCD2023 per presentare nuovamente la propria infrastruttura ed invitare tutti colleghi interessati a farne uso e a formulare idee e proposte di collaborazione per un possibile ampliamento del corpus.

<sup>4</sup> Beta: [https://github.com/DanilSko/udracor/blob/ezdramaformat/ezdramaparser\\_version\\_20220928\\_prose\\_verse\\_switch.ipynb](https://github.com/DanilSko/udracor/blob/ezdramaformat/ezdramaparser_version_20220928_prose_verse_switch.ipynb)

<sup>5</sup> <http://www.bibliotecaitaliana.it/catalogo>

<sup>6</sup> Disponibile su portali quali [opera-guide.ch](http://opera-guide.ch), [opera.stanford.edu](http://opera.stanford.edu), [librettidopera.it](http://librettidopera.it), [operalib.eu](http://operalib.eu), oltre che nei siti collegati al *Progetto Metastasio*, attivo presso l’Università di Padova ([progettometastasio.it](http://progettometastasio.it), [apostolozeno.it](http://apostolozeno.it), [carlogoldoni.it](http://carlogoldoni.it), [variantiallopera.it](http://variantiallopera.it)). Attualmente è in sviluppo una pipeline per lo *scraping* di tali testi e la successiva conversione; il progetto è aperto a eventuali collaborazioni esterne.



## 4. RINGRAZIAMENTI

DraCor ha ricevuto finanziamenti tramite il programma Horizon 2020 dell'Unione Europea (*grant agreement* n. 101004984) nell'ambito del progetto CLS-INFRA (<https://clsinfra.io>). L'elenco completo dei contributori al progetto DraCor è disponibile al seguente link: <https://dracor.org/doc/credits>.

## BIBLIOGRAFIA

- [1] Börner, Ingo, Frank Fischer, Luca Giovannini, Christopher Lu, Carsten Milling, Daniil Skorinkin, Henny Sluyter-Gäthje e Peer Trilcke. «Onboard onto DraCor: Prototyping Workflows to Homogenise Dramatic Corpora». In: DHd2023: Open Humanities, Open Culture. Book of Abstracts. <https://doi.org/10.5281/zenodo.7688632>
- [2] Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtel, Christopher Kittel, Carsten Milling e Peer Trilcke. «Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama». In DH2019: Complexities. Book of Abstracts. <https://doi.org/10.5281/ZENODO.4284002>
- [3] Fischer, Frank, Gilles Dazord, Mathias Göbel, Christopher Kittel, e Peer Trilcke. «Le drame comme réseau de relations. Une application de l'analyse automatisée pour l'histoire littéraire du théâtre». *Revue d'Historiographie du Théâtre* 4 (2017). <https://hal.science/hal-01811799>
- [4] Fischer, Frank, Mathias Göbel, Dario Kampkaspar, Christopher Kittel e Peer Trilcke. «Network Dynamics, Plot Analysis. Approaching the Progressive Structuration of Literary Texts». In: DH2017: Access/accès. Book of Abstracts. <https://dh2017.adho.org/abstracts/071/071.pdf>
- [5] Moretti, Franco. «Conjectures on World Literature». *New Left Review* 1 (2000): 54-66.
- [6] Mueller, Martin. «Shakespeare His Contemporaries: Collaborative Curation and Exploration of Early Modern Drama in a Digital Environment». *Digital Humanities Quarterly* 8, 3 (2014). <http://www.digitalhumanities.org/dhq/vol/8/3/000183/000183.html>
- [7] Quondam, Amedeo. «Memorie per una storia dell'italianistica digitale: "Biblioteca Italiana"». *Griseldaonline* 20, 2 (2021): 137-47. <https://doi.org/10.6092/issn.1721-4777/12360>
- [8] Trilcke, Peer, Christopher Kittel, Nils Reiter, Daria Maximova e Frank Fischer. «Opening the Stage: A Quantitative Look at Stage Directions in German Drama». In: DH2020: carrefours/intersections. Book of Abstracts. [https://dh2020.adho.org/wp-content/uploads/2020/07/337\\_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html](https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html).

# Metodologie di indagine per una nuova piattaforma: Digital Scholarly Editions Platform and Aligned Translation

Hansmichael Hohenegger (Principal Investigator)<sup>1</sup>, Federico Boschetti<sup>2</sup>, Iacopo Chiaravalli<sup>3</sup>, Fabio Ciotti<sup>4</sup>, Eleonora De Longis<sup>5</sup>, Angelo Mario Del Grosso<sup>6</sup>, Tiziana Mancinelli<sup>7</sup>, Federico Meschini<sup>8</sup>

<sup>1</sup>Istituto Italiano di Studi Germanici, Italia – hohenegger@studigermanici.it

<sup>2</sup>Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILC), Italia – federico.boschetti@ilc.cnr.it

<sup>3</sup>Istituto Italiano di Studi Germanici, Italia – chiaravalli@studigermanici.it

<sup>4</sup>Università degli Studi di Roma Tor Vergata, Italia – fabio.ciotti@uniroma2.it

<sup>5</sup>Istituto Italiano di Studi Germanici, Italia – delongis@studigermanici.it

<sup>6</sup>Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILC), Italia – angelo.delgrosso@ilc.cnr.it

<sup>7</sup>Istituto Italiano di Studi Germanici, Italia – tiziana.mancinelli@gmail.com

<sup>8</sup>Università degli Studi della Tuscia, Italia – fmeschini@unitus.it

## ABSTRACT

La Digital Scholarly Editions Platform and Aligned Translations (DiScEPT, Progetto Straordinario FOE finanziato dal MUR: 2022 – 2024) è un ecosistema digitale sostenibile che integra strumenti di codifica testuale e visuale in un'interfaccia coerente che consente sia la costituzione e curatela di Digital Scholarly Editions (DSE) sia la loro pubblicazione e fruizione. A tal fine sono stati sviluppati protocolli di ricerca empirica (qualitativa e quantitativa). Qui presentiamo un sistema di interviste strutturate da somministrare alle figure professionali che operano nel campo delle DSE. Il fine è raccogliere le loro esigenze e i requisiti per progettare una piattaforma che ad esse risponda.

## PAROLE CHIAVE

Survey questionnaires, Needs and Requirements Gathering, Digital Scholarly Editions, Aligned Translations.

## 1. INTRODUZIONE

Il progetto DiScEPT sviluppa un ecosistema digitale sostenibile per la produzione e la pubblicazione di edizioni scientifiche digitali (DSE), integrando servizi di ausilio all'editing con strumenti che possano affiancare varie versioni di un testo o di corpora interi allineando traduzioni in varie lingue. Per far tesoro delle esperienze degli ultimi decenni [1] [2] [3] [4] [5] [6], si deve tener conto delle diverse, e talvolta contrastanti, esigenze di coloro che producono edizioni di testi digitali, così come di coloro che ne fruiscono. Lo sviluppo dell'architettura della piattaforma è quindi affiancato sin da subito da protocolli di indagine sui soggetti interessati (filologi tradizionali e digitali, editori digitali, *digital humanist*, traduttori, grafici, etc.). Il sistema di interviste, o *survey questionnaires*, è complementare alle altre modalità per raccogliere esigenze e necessità alle quali vuole rispondere il progetto: focus group e seminari [7], interviste a risposta aperta per l'analisi qualitativa [8] e l'osservazione partecipativa. In occasione del XII Convegno annuale AIUCD 2023 presentiamo qui lo schema di progettazione dell'architettura della piattaforma insieme a un questionario strutturato rivolto agli studiosi presenti al quale potete accedere tramite QR code.

## 2. ARCHITETTURA DELLA PIATTAFORMA

L'idea alla base della piattaforma DiScEPT emerge dalla necessità di fare tesoro di numerose sperimentazioni, teorie e pratiche pluridecennali che non riescono, però, a concretizzarsi in modelli di DSE che vadano oltre uno strumento di pubblicazione digitale pensato per la singola edizione pur raffinatissima. Il punto di forza di DiScEPT non è tanto la novità nell'elaborazione dei dati o nello svolgimento delle attività multiple connesse alla produzione editoriale e alla loro fruizione, quanto *l'integrazione di strumenti e servizi già disponibili e ben noti a editori e studiosi in una interfaccia coerente e di facile utilizzo*.

Obiettivo di DiScEPT è sviluppare un ecosistema che, oltre a rappresentare un'applicazione di produzione di contenuti, sia pensato anche per la loro fruizione. Proprio per questo l'ecosistema sarà distinto in due ambienti principali a seconda delle modalità di utilizzo.

1. **Back-end:** *virtual workbench* flessibile e modulare, ma allo stesso tempo capace di gestire dati strutturati ed esporre standard (sia di formato sia di processo) consolidati nella comunità della *digital philology*. La prima scelta strategica in una piattaforma basata sui dati è, infatti, quella di definire il *data model* e il conseguente formato di rappresentazione, che, nel campo delle edizioni digitali, è stato ormai individuato dalla comunità scientifica di riferimento nel formato di codifica dei dati XML – e in particolare nello schema di codifica predisposto dalla TEI – insieme a LOD e alle tecnologie semantiche. Il progetto intende adottare questo formato come scelta prioritaria, ma anche di sperimentare approcci innovativi come i workflow basati sui Domain Specific Languages (DSL), sviluppati dal gruppo di ricerca del CoPhiLab del CNR-ILC [9] [10]. Insieme alla piattaforma di editing, saranno integrati i seguenti moduli operativi: a) digitalizzazione delle fonti a partire dai testimoni cartacei; b) OCR e HTR basato su reti e Machine Learning; c) collazione automatica; d) piattaforma di annotazione multilayer del testo e di associazione di metadati complessi; e) sistema di visualizzazione, presentazione e pubblicazione; f) sistema di supporto alla traduzione con tecniche IA neurali; g) Accesso a lessici e vocabolari multilingua; g) sistema di accesso ad archivi digitali testuali multilingue per il supporto allo studio della tradizione e l'analisi intertestuale.
2. **Front-end:** interfaccia virtuale che permetta la pubblicazione e consultazione delle edizioni attraverso, per esempio, il *framework* IIF, per permettere l'accesso, l'interoperabilità e il riuso degli oggetti digitali. La componente di pubblicazione integrerà un modulo di gestione dei diritti di accesso ai materiali pubblicati. Inoltre, la piattaforma renderà possibile generare versioni camera ready ad alta qualità tipografica. È previsto, infatti, che si debba poter pensare a una pubblicazione anche in formato cartaceo.

### 3. RUOLO DELLE INTERVISTE E LORO STRUTTURAZIONE

Il progetto DiScEPT, in quanto esito delle esperienze di editoria digitale e delle riflessioni in merito degli ultimi decenni, necessita di una continuata raccolta di dati. Non si tratta tanto di marcare una discontinuità con le esperienze precedenti, quanto sfruttare i risultati di un'attenta *gap analysis* ottenuta grazie al confronto tra i modi di vedere del progettista digitale, del filologo, del *digital humanist*, del traduttore, e degli altri *stakeholder* (lettori, studiosi, docenti, editori commerciali ecc.).

Il sistema di interviste strutturate vuole rispondere a questa esigenza. È stato indispensabile porsi il problema di che cosa sia un questionario, di come agisca e di come sia possibile strutturarne in modo tale che, da un lato, esso sia adatto ai soggetti di riferimento, ma che, dall'altro, quegli stessi soggetti abbiano la possibilità di intervenire sulle modalità stesse dell'indagine, sul modello dell'*interactive and dynamic data analysis*. La prima decisione presa è stata quella di assumere il modello dell'intervista strutturata come strumento principale. La possibilità di ulteriori confronti e discussioni è stata invece demandata al rapporto con figure autorevoli in settori specifici (a ciò si aggiunge poi un'indagine di tipo sperimentale che, grazie all'osservazione partecipativa di matrice antropologica, punta a discernere gli elementi fondamentali dell'attività psicomotoria dei professionisti del testo). I diversi tentativi di indagine simili sono stati raccolti e utilizzati come metro di confronto per capire come fossero migliorabili e adattabili alle nostre esigenze. Ciò ha consentito la costruzione di un sistema scalare di interviste diverse, divise, in primo luogo, in base alle due principali aree operative del software (si veda punto 2). All'intero di questa macrodivisione, il questionario è stato poi distinto fra le categorie professionali coinvolte nella produzione degli oggetti editoriali (testi digitali, metadati, immagini, archivi / repository, software, piattaforme, ecc.) o ai soggetti coinvolti nella loro produzione o fruizione (filologi, editori digitali e non, traduttori e grafici, lettori e editori commerciali). A ognuna di queste figure è stata associata una sequenza di quesiti a risposta chiusa. Tutte le diverse tipologie di sequenze si chiudono con una sezione volta a consentire al soggetto intervistato una valutazione dell'intervista stessa, così da risultare correggibile e modificabile nel corso del tempo in base all'allargarsi e all'approfondirsi della ricerca. La somministrazione delle diverse interviste è stata adattata ai diversi contesti ed è stata svolta attraverso modalità multiple e modulate che vanno dalla compilazione individuale all'intervista guidata da parte dell'intervistatore.

### BIBLIOGRAFIA

- [1] McGann, Jerome. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*, Harvard University Press, 2014.
- [2] Meschini, Federico. *Oltre il libro. Forme di testualità e digital humanities*, Editrice Bibliografica, 2020.
- [3] Driscoll, Matthew James, Pierazzo, Elena. *Digital Scholarly Editing: Theories and Practices*, Open Book Publishers, 2016.
- [4] Schillingsburg, Peter L. *Scholarly Editing in the Computer Age. Theory and Practice*, University of Michigan Press, 1996.
- [5] Tomasi, Francesca, *Organizzare la conoscenza: digital humanities e web semantico*, Editrice Bibliografica, 2022.



- [6] Pierazzo Elena, Mancinelli Tiziana, *Che cos'è un'edizione scientifica digitale*, Carocci, 2020
- [7] Acocella, Ivana. «L'uso dei focus group nella ricerca sociale: vantaggi e svantaggi». *Quaderni di sociologia* 37 (2005): 63 – 81.
- [8] Marradi, Alberto. *Tutti redigono questionari. Ma è davvero così facile?*. Franco Angeli, 2019.
- [9] Boschetti, Federico, Del Grosso, Angelo Mario. «L'annotazione dei testi storico-letterari al tempo dei social media». *Italica Wratislaviensia* 11 (2020): 65 – 69.
- [10] Del Gratta Riccardo, Del Grosso Angelo Mario, Zenzaro Simone, Boschetti Federico, Bambaci Luigi. «La Filologia come sistema dinamico». *Umanistica Digitale* 13 (2022): 1-20.

# Making Korean Perseus Digital Library: The Case Study of ‘Dokdo Digital Library’

Jaewon Ahn<sup>1</sup>, Hanuri Son<sup>2</sup>, Kwangho Kim<sup>3</sup>, Seokchan Yun<sup>4</sup>

<sup>1</sup> Institute of Greco-Roman Studies, Republic of Korea – numeniu@snu.ac.kr

<sup>2</sup> Interdisciplinary Program in Classical Studies, Seoul National University, Republic of Korea – thsgksnfl@snu.ac.kr

<sup>3</sup> Interdisciplinary Program in Classical Studies, Seoul National University, Republic of Korea – bdpq139@snu.ac.kr

<sup>4</sup> Interdisciplinary Program in Classical Studies, Seoul National University, Republic of Korea – yschespero@snu.ac.kr

## ABSTRACT

While the libraries of major countries in Europe and North America have actively engaged in digitizing their manuscripts, so-called Digital Libraries in South Korea have been mere archives of photoprints or digitized documents. This gap is largely due to the misinterpretation of their own core values: Digitization and Reliability. The digital libraries in Korea have not properly digitized the source materials and have not provided reliable texts from the perspective of classical philology. This current state has been shown that the renovation of Korean digital libraries is an urgent and a quintessential task.

This article aims to identify new editorial guidelines and production methods to create digital editions of Modern Korean texts, with the case study of Dokdo Digital Library (i.e. DDL). DDL proposes texts based on the principles of reliability, standardization, accessibility, usability, connectivity, and extendibility. DDL goes beyond the existing limitations of printed media and is realized in digital space. It not only provides users with a dynamic experience of enjoying materials but also builds dynamic networks between various texts that cannot be captured in printed media. Consequently, It will produce a variety of new research topics, as well as lay the foundation for an academically reliable Korean lexicon, TLC(*Thesaurus Linguae Coreanae*). This article will introduce plans for the DDL project and show specific aspects of DDL. The first product of the DDL project, *Tongyangpyŏnghwaron*, which was written by Korean independence activist An Chung-gŭn, is an excellent example. We hope that DDL will become a cornerstone for Korean philology and digital humanities research.

## KEYWORDS

DDL(Dokdo Digital Library); Digital Humanities; Korean philology; Modern Korean texts

## 1. INTRODUCTION

So-called ‘Digital Libraries’ in Republic of Korea have shown misinterpretation of their own core values: Digitization and Reliability. Most of the existing Digital Libraries have ended up uploading mere photoprints or digitized documents. The uploaded texts, therefore, left behind all other fundamental information including original texts, digitized copies, references, annotations, archival information and so forth. They are also known to be lacking their efforts on organizing these various sets of information in a systematic manner.

The reliability of their collection has also been problematic, as most of the texts are not in the form of critical edition. Most of the existing digital texts, as shown above, do not incorporate basic information such as annotations, their editors and translators. Especially, early modern Korean texts have been proved more problematic as the information on their modifications regarding variant letters, punctuations, word orders and lacunas has not frequently been specified.

Those major vulnerabilities of current Korean Digital Libraries are largely due to the absence of unified editing principles in this country. This does not mean that Korea lacked history of philology. In fact, Koreans have edited *Sŏnbon* (善本, *editio bona*), i.e. the most academically valuable edition, and have translated *Sŏnbon* into modern Korean language for a long time. However, this does not indicate that the Korean tradition of text selection and translation does not have its limitation; *Sŏnbon* is not a clear concept and the tradition lacks critical apparatus and text tracking.

Overcoming those limitations and providing philologically organized digital texts are major, utmost concerns of Dokdo Digital Library Project. DDL is a publicly funded digital library that aims to collect, organize, and create a digital critical edition of modern Korean texts and ultimately provide the new experience that has been unattainable in printed media, and the establishment of links between the resources. DDL especially concentrates on texts written between the late 19th and early 20th centuries, which provide insight into the transition from pre-modern Korean to modern Korean language and the

formation of Korean civilization. Following chapters will specify its prospects by clarifying the basic principles, basic templates, example studies of DDL.

## 2. PRINCIPLES OF DDL

As stated above, DDL Seeks to overcome the existing limitations of Korean Digital Libraries. Setting Perseus Project of Tufts University as a model, DDL team has established three following principles.

The first principle is reliability and standardization. The texts provided in DDL are edited as the critical editions following the philological methodology. This would provide reliability by specifying their editors, translators and annotators. The apparatus criticus and editing signs observe international standards and own developed signs also can be provided to describe Korean texts and Classical Chinese texts.

The second principle is accessibility and usability. DDL is operated under the financial support of *Tokodosögwanch'in'gudül* (독도도서관친구들, *Friends of the Public Library in Korea*), a NGO working for the nationwide provision of public libraries. As such, DDL Project aims to upload all of its archival possessions in an open online platform, enabling citizens to access the material regardless of their geographical, economic, and cultural constraints. This opening of the information entails user-friendly improvement of designs and composition for the usability of professional researchers as well as the public.

The third principle is connectivity and extendibility. DDL aims at constructing information networks on the basis of the intertextuality among its archival possessions. This would lead readers to other related texts, providing multi-textual view on the matter of interest. For instance, with current archival possessions regarding Korean modernity, DDL can operate as a powerful source material by giving out multi-dimensional viewpoints on modern Korean ideology, society, and language. Especially, its possession being classified with its ages and authors, DDL can help linguists trace the semantic and syntactic transition that had occurred in the history of Korean language.

The network would also incorporate information from external sources, including digitized images of objects, places, inscriptions, and printed pages, geographic information, and other digital representations of objects and spaces, as provided in Perseus Project. This would provide the experience unattainable in the printed format. The external sources include other text archives. For instance, the ongoing digital edition of *Yönsölböppang* (演說法方, *The way to make speeches*) would provide hyperlinks for the quotations of various orators including Cicero and Demosthenes, by linking the quotation to its original texts in other digital archives such as Perseus Project.

With these three principles, DDL will ultimately provide model for other digital archives in Korea and universal access to modern Korean literature.

## 3. THE TEMPLATE OF DDL

The following are the basic features that DDL aims to provide.

Features	Detailed functions
Text-related features	Original text, Translation text, Apparatus criticus, Commentary
Search-related features	Author search, Text search, Passage search, Word search
Lexicon-related features	Word list, Usage, Pronunciation, Definition, Grammar
Reference-related features	Source list, Research list, Map list, Image list

Table 1. Main Features.

Firstly, Text-related features can be implemented with ease, for those features mainly focus on the "digitization" of existing printed critical editions. For now, placing emphasis on public access, the entire web page and the text have been created in HTML format. But we will plan to switch them into TEI-compliant XML texts, concentrating on original texts and translations to automate data processing and ensure scalability.

Search-related features are neither necessary nor beneficial right now, as the amount of accumulated text is still quite small. However, once the volume of texts is sufficient, they will function as important features in providing accessibility and convenience to users. Among the Lexicon-related features, Word list, Usage, and Pronunciation were implemented first. Considering the reality Korean dictionaries have been placed in, Definitions and Grammar are not yet available. As most of the existing Korean dictionaries were not summarized according to TLL principles, they are at best a collection of sentences and excerpts that is made by the author on the basis of his own linguistic intuition. Those are not academic dictionaries in a strict sense. In short, there are very few annotated dictionaries available for DDL, especially for modern

Korean. It is comparable to the fact that Perseus was able to utilize lexicons that had already accumulated enough usage based on a long academic tradition.

Given this situation, we do not expect DDL to work as a source for lexicons. Rather, we hope that the data collected through DDL can be the primary source for creating Korean Lexicon, TLC(Thesaurus Linguae Coreanae). Some prospects of TLC would be discussed in the final chapter. For reference-related features, both texts and relevant research materials must be collected first, so that we can reach a suitable point where those data can be catalogued and made available to the public. We think that the publication of *Hwasŏngdonjŏn*(華盛頓傳, *The Biography of George Washington*), which is currently in preparation for publication, is expected to be a turning point, as it contains a large amount of secondary sources such as geographic information, battle maps, archaeological data, and biographical information, usw. Below are sample footages that demonstrate the features currently implemented in DDL Web.



Figure 1. Main page.



Figure 2. Text List.

Figure 3. Text page.

Figure 4. Interpretation page.

#### 4. THE CASE STUDY OF TONGYANGP'YŎNGHWARON (東洋平和論, THE DECLARATION FOR THE PEACE OF ASIA)

The case study on the first achievement of the DDP project, *Tongyangp'yŏnghwaron* (hereafter DPR), would provide an outlook of the ongoing edition and digitization of the project. DPR was published with 9 following steps: 1. Overall

investigation on the work, 2. Literature tracking, 3. Textual criticism, 4. Comparative verification of the text, 5. Organization and revision of the text, 6. Translation of the text, 7. Annotation of the text, 8. Collection of modern terminology, 9. Digitization.

1. An Chung-kūn (安重根, 1879-1910, here after An), the author of DPR, was a Korean independence activist against Japanese imperialism. He was also a nationalist and a pan-Asianist. With his passion towards Korean independence, he assassinated Itō Hirobumi (伊藤博文, 1841–1909), ex-Japanese prime minister and ex-Korean resident-general on the October 26<sup>th</sup>, 1909. Due to the assassination, he was arrested by Japanese army and imprisoned, sentenced to death in Japanese Court. He composed this treatise until his execution in March 1910, leaving DPR incomplete.
2. As the table of contents shows, An tried to compose DPR in five chapters: *Sō* (序, Introduction), *Chōn'gam* (前鑑, The lessons from the past), *Poksōn*, (伏線, The groundwork and context), *Hyōnsang* (現狀, The current state of affairs), and *Mundap* (問答, Questions and answers).<sup>1</sup> Among those, only the first two chapters were completed. However, despite its incompleteness, DPR still gives many clues to his ambitious plan of Korean independence and Pan-Asiatic Peace. He strongly insists that the Peace is not to be achieved through the imperialist rule of peace and force, rather through *Kongjon*(共存, co-existence), a term he used to describe the state of international acknowledgement of equal sovereignty, and *Kongyōng* (共榮, co-prosperity), a term for not attempting foreign invasion and pursuing prosperity. Therefore, His novel prospect, although incomplete, played a significant role in Korean independence movement and the formation of Korean identity.
3. The autographic manuscript of An is yet to be spotted. Currently there are one manuscript, three photoprints, and one edition identified as DPR. The only manuscript is stored in Japanese National Diet Library(国立国会図書館, hereafter NDL). The three photoprints are each published by Modern Japanese Political History Materials Room(憲政資料室, hereafter PHMR) of NDL, Yun Pyōng-sōk(윤병석, hereafter Yun) in 2010, and An Chung-kūn Peace Research Institute(안중근평화연구원, hereafter APRI). The edition was published in 2016 by APRI.
4. The NDL Manuscript was known to be the first-hand manuscript copy of An's Japanese prison guard. The PHMR and Yun Photoprint is the photoprint of the NDL Manuscript, and the APRI photoprint is known to be the reproduction of the PHMR photoprint. The APRI edition was based on the PHMR photoprint, and have contained the PHMR photoprint itself as an appendix.
5. As the NDL did not allow lending out their possessions, the PHMR photoprint became the basis of the new edition. The only existing edition by APRI lacked consistent editing principles and had a number of errors. However, due to its uniqueness, the APRI edition has become one of the objects of comparative research.
6. The comparative researches on the existing copies found out two major types of their errors. First, existing copies have arbitrarily erased or added punctuations. Secondly, typos are frequently spotted. The typos could be again classified with the causes of their occurrences: namely, Handwriting errors, Deciphering errors, Pinyin and Simplified Chinese input errors, Synonymous letter input errors, Omission errors, Addition errors, Corrupted word order errors, and Errors in spacing. These errors and their corrections are recorded in the critical apparatus. The corrections followed international standard for the most part, but have included new several apparatuses for the Korean texts, such as '々', noting the repetition of the preceding letter.
7. The text is organized into three parts: preface, body paragraphs, and apparatus criticus. In the preface, The contents regarding literature tracking, textual criticism, and the introduction to each copy was described as above, and introduced the Sigla used in the text. The body paragraphs included paragraphs corresponding to the pages of the PHMR photoprints in the margin, and the numbers of each paragraph (1a, 1b, 2a ...), and the numbers of each line. (1, 2, 3...) The Critical Apparatuses were located underneath the body paragraphs and gave out the correction and editing history, even improvements of the text if necessary.
8. Before moving on to the translation, DDL team has examined 6 existing translations of DPR. Most of the translations were mere paraphrases of Ch'oe Sun-hŭi(최순희, 1979). Among those, the translation of Ch'oe Kūn-tōk (최근덕, 1979) turned out to be the most trustworthy version, and Yun's photoprint and translation was a paraphrase of Ch'oe's version. The biggest problem of the existing translations was that the literature evidences of the translation was not given. This has caused various errors in their translations. In contrast, DDP team has

---

<sup>1</sup> English translations are from Ahn (2022)

translated on the basis of critical edition, and left annotations to compare with the significant errors that other versions have made.

9. Annotations included introductions to major figures and regions, historical and temporal backgrounds, and quotations to other authors (e.g. Mencius), and have provided interpretations of the significant phrases and their academic evidences.
10. Modern terminologies were collected on the completed edition. Modern terminologies included vocabulary not found in preceding ages, which includes proper names for people and regions, newly translated words, terminologies regarding the new technologies, and the words that have changed its meaning in modern era. The vocabulary includes *Miguk*(美國, The United States), *Ŭiguk*(義國, Italia), *Taet'ongnyŏng*(大統領, President), *Ch'imsujŏng*(浸水艇, submarine) and *Taehandongnip* (大韓獨立, Korean Independence) as its examples. These would later be the foundation for the Korean Lexicon based on DDL.
11. For the final step, the critical edition, translation and annotations were uploaded onto the DDL website. Secondary sources such as books and research essays are provided onto each paragraph for the sake of additional research. Digitized images, geographic information, and other digital materials are provided as well. DDL team has also created lists for modern terminology, noting usages from other texts on DDL website, secondary sources and digital materials. This would make reading experience comprehensive by providing links between the body text and the terminology lists. Listed below are its examples.

The screenshot shows the DDL website interface. At the top left, there is a logo for 'DDL' and the title '근대 용어' (Modern Vocabulary). A navigation menu on the left lists various terms like '덕국', '대련', '대만', '대한독립', '동양순양함대', '동양인종', '동양평화', and '동양평화론'. The main content area is titled '대한독립 大韓獨立' and contains two sections: '동양평화론' (East Asian Peace Theory) with a reference to '1b 大韓獨立鞏固云。' and '안응칠역사' (An Ung-chil's History) with several references including '하2 當大韓獨立之日。', '하13 大韓獨立鞏固之意。', '하42 大韓獨立鞏固云。', '하87 大韓獨立萬歲。', and '하99 大韓獨立。'. The right sidebar has two sections: '참고문헌' (References) listing books like '대한독립운동사' and '안중근의사', and '관련자료' (Related Materials) listing a book '안중근의사의 혈서 태극기' and a website 'EBS\_정물\_0545'. At the bottom of the sidebar, there is a photograph of a wall with the large Korean characters '대한독립' (Great Korean Independence) and a circular emblem.

Figure 5. Word List.

## 5. KOREAN PROSPECT

For the upcoming projects, DDL is currently working on three texts: *Anŭngch'iryŏksa* (安應七歷史, Autobiography of An Eungchil), *Hwasŏngdonjŏn* (華盛頓傳, Biography of George Washington), and *Yŏnsŏlbŏppang*.

1. *Anŭngch'iryŏksa* is the autobiography of An Chung-kŭn, the author of DRP. He wrote the book in prison between the 13th December 1909 and the 15<sup>th</sup> March 1910, 12 days before his execution. Like DRP, his autographic handwriting is yet to be discovered, and currently accessible copies are all based on Japanese manuscripts of the day.

*Aneungchilyeoksa* provides another viewpoint of An Chung-kŭn ideology. The core concept of his ideology is 'In(仁)', meaning love in that he has insisted to confront the evil with 'In'. This 'In' works as the basis of his 'Gongjon' and 'Gongyeong' theory in that it insists tolerance towards the others.

2. *Hwasŏngdonjŏn* is the biography of George Washington written by Lee Haejo (李海朝, 1869-1927), a Korean new-style novel writer. Lee Haejo had looked up to the biographies on other languages, including the 1903 Chinese version



*huashengdun*(華盛頓) by Dingjin(丁錦) and 1900 Japanese version *Washington*(華聖頓) by Fukuyama Yoshiharu(福山義春), while writing his book in Korean-Classical Chinese hybrid. It is known that 1900 Japanese version has acted as the basis of the other two, and *Hwasōngdonjōn* is mostly composed of the translation from the 1903 Chinese versions. The three versions have introduced the background information of G. Washington in detail, and have described his achievements from the French and Indian War towards the end of American Independence War. The books did not miss related information, including geography, names, and individual battles and battleships, helping the readers of three nations understand G. Washington and the United States.

The upcoming DDL version is composed of the critical editions of *Hwasōngdonjōn*, 1903 Chinese version, and their Korean translations. Comparative annotations with 1900 Japanese versions are provided in important phrases, enabling comparisons of the reception on G. Washington among three nations. The DDL version would also be the important source of tracking geography and historical figures in early American history.

In the Second volume of *Aneungchilyeoksa*, An Chung-kūn has complimented G. Washington on his role in American Independence. This reference might have influenced the decision of Japanese government officials to include *Hwasōngdonjōn* in its *Index Librorum Prohibitorum*.

3. *Yōnsōlbōppang* is the guide to the rhetoric by An Gukseon(安國善, 1878-1926), a Korean writer. He has described the history of speech, the manner of orators, the training methods on the speech, and the closing method of the speech by introducing several notable speeches such as those of Cicero and Demosthenes as examples. This textbook might have been used as the text book in Donmyeongeuisuk(敦明義塾) as An Gukseon was the politics professor of the school.

Speeches and panel discussions then played an important role in overcoming the delivery problems confronted with printed materials which are meant to enlighten the public with the sense of modernity. *Yōnsōlbōppang* is also important in present sense as it not only helps research regarding the Korean culture of speech and debate, but also research regarding the reception of Western intelligence with its reference to Western thinkers.

These upcoming texts would contribute to the modern terminology data in the DDL website. This would contribute to the further project of Korean Digital Lexicon, where the data regarding the usages of the archive are interactively and systematically linked with each other to produce valuable data of the vocabulary all the way from its reception to the grammatical information.

These accumulations would contribute to reverse the current status of Korean philology itself. Europeans, from the very dawn of publication, have developed rich philological tradition of Stemma Codicum and critical editions. In contrast, Koreans, in spite of their rich history of publications known for *Jikji*(直指) and *Tripitaka Koreana*(八萬大藏經), have not yet developed such tradition. The digitization can be the key of reversing this backwardness by simultaneously creating the digitized copy and the critical edition. Current Korean perception towards the digitization as the provision of data can be the obstacle, but the space for the persuasion still remains, as these newly digitized texts would have huge impacts on humanities by providing the source of education and influence. This requires a long period of time, explaining the reason why DDL project has chosen *Toktodosōgwanch'in'gudūl* as its patron. While the government funded project in Korea is still swayed by short-term political interest, *Toktodosōgwanch'in'gudūl*, with its 'one-Euro-a-month program' participated by 3000 and more, has promised a prolonged support to the program. Supported as such by many, the DDL project will be one of cornerstones for the Korean Language Civilization.

## REFERENCES

[1] Ahn, Junggeun. *The Declaration for the Peace of Asia : Critical Edition*. Edited by Munseok Kwak, Eunsuk Kim, Taeju Kim, Jaewon Ahn. Seoul: Dokdo Friends of Public Library in Korea, 2019. [안중근. 동양평화론 : 비판정본. 광문석, 김은숙, 김태주, 안재원, 윤재성 역. 서울: 독도도서관친구들, 2019.]

[2] Ahn, Junggeun. *The Autobiography of Ahn Eung Chil : Critical Edition*. Edited by Eunsuk Kim, Taeju Kim, Jaewon Ahn, Jaeseong Yun. Seoul: Dokdo Friends of Public Library in Korea, 2020. [안중근. 안응칠 역사 : 비판정본. 김은숙, 김태주, 안재원, 윤재성 역. 서울: 독도도서관친구들, 2020.]

[3] Lee, Haejo. *The Biography of George Washington : Critical Edition*. Edited by Eunsuk Kim, Taeju Kim, Hanuri Son, Jaewon Ahn. Seoul: Dokdo Friends of Public Library in Korea, forthcoming. [이해조. 화성돈전: 비판정본. 김은숙, 김태주, 손하누리, 안재원 역. 서울: 독도도서관 친구들, 출판 중.]

[4] Ahn, Jaewon. *Ad fontes*. Seoul: Nonhyeong, 2019. [안재원. 원천으로 가는 길. 서울: 논형, 2019.]

- [5] Ahn, Jaewon. "Does Pax Mean Truly Peace? Focusing on The Declaration for the Peace of Asia of Ahn Junggeun (안중근)" In *Empire and Politics in the Eastern and Western Civilizations: Searching for a 'Respublica Romanosinica*, edited by Andrea Balbo, Jaewon Ahn and Kihoon Kim, 239-254. Berlin, Boston: De Gruyter, 2022. <https://doi.org/10.1515/9783110731590-015>
- [6] Choi Chaegi. *Introduction to Arrangement of Classical Books*. Daejeon: Hangminmunhwasa, 2011. [최채기. *고전적정리입문*. 대전: 학민문화사, 2011.]
- [7] Hwang Yeongnyeon. *Introduction to Arrangement of Classics*. Institute for the Translation of Korean Classics [황영년. *고적정리개론*. 서울: 한국고전번역원, 2013.]
- [8] Kim Hyeon. *Introduction to Digital Humanities*. Seoul: HUEBOOKs, 2016. [김현. *디지털 인문학 입문*. 서울: HUEBOOKs, 2016.]
- [9] Lee Jungwon. *The Future of Humanities in the Digital Era*. Seoul: Pureunyeoksa, 2017. [이중원. *디지털 시대 인문학(人文學)의 미래*. 서울: 푸른역사, 2017.]
- [10] Ahn, Jaewon. "Eine historisch-methodische Bemerkung zum Thesaurus Linguae Latinae" *The Journal of Greco-Roman Studies* 25. (2006): 359-402. [안재원. "라틴어 대사전에 대하여." *서양고전학연구*, no. 25. (2006): 359-402..]
- [11] Ahn, Jaewon. "Some Suggestions for Critical Edition of Korean Classics" *Korean Studies Quarterly* 35, no.3 (2012): 31-57. <https://doi.org/10.25024/ksq.35.3.201209.31> [안재원. "왜 '정본'인가." *한국학* 35, no.3 (2012): 31-57.]

# Constructing an Old English WordNet: The Case of *Guilt*

Anas Fahad Khan<sup>1</sup>, Michele Cavallaro<sup>2</sup>, Rafael Cruz González<sup>3</sup>, Javier Díaz-Vera<sup>4</sup>,  
Francesca Frontini<sup>5</sup>, Francisco Javier Minaya Gómez<sup>6</sup>

<sup>1</sup> Istituto di Linguistica Computazionale ‘Antonio Zampolli’ (CNR-ILC), Italy – fahad.khan@ilc.cnr.it

<sup>2</sup> University of Siena, Italy – michele.cavallaro@student.unisi.it

<sup>3</sup> Universidad de Castilla-La Mancha, Spain – rafael.cruz@uclm.es

<sup>4</sup> Universidad de Castilla-La Mancha, Spain - javierenrique.diaz@uclm.es

<sup>5</sup> Istituto di Linguistica Computazionale ‘Antonio Zampolli’ (CNR-ILC), Italy – francesca.frontini@ilc.cnr.it

<sup>6</sup> Universidad de Castilla-La Mancha, Spain – francisco.minaya@uclm.es

## ABSTRACT

In this paper, we look at the manual construction of a lexicon of emotion terms in Old English organised as a wordnet lexicon and based on a pre-existing dataset which categorises emotion terms on the basis of cognitive criteria. This is part of a wider initiative to construct a wordnet for Old English which includes both manually and (semi-)automatically constructed parts. We will describe our workflow which foresees the use of lemmas, senses and definitions taken from and in many cases adapted from a legacy Old English lexical resource, ClarkHall’s AngloSaxon Dictionary, and organised into synsets by matching the definitions of OE terms with the synset definitions found in the Open English WordNet. We argue that this is a useful way of disseminating research in lexical semantics and philology taking as an example the semantic field of GUILT. In our poster presentation we will give a visual description of the whole field; in the present submission we only focus on the most ‘core’ nominal synsets relating to GUILT.

## KEYWORDS

Old English; wordnet; lexicography; guilt

## 1. INTRODUCTION

The work described in this submission is part of a wider project which aims towards the creation of a wordnet for the Old English<sup>1</sup> language (OE), the *Old English WordNet* (OldEWN), a project which is described in detail in previous works [3,4]. The OldEWN is being constructed semi-automatically<sup>2</sup> on the basis of the information contained in a public domain Old English dictionary, J. R. Clark-Hall’s *Concise Anglo-Saxon Dictionary* (1916) (CH) [1]. The entries in this latter work are short and easy to process; something which distinguishes it both from the earlier and much more comprehensive Old English dictionary Bosworth and Toller’s *Anglo-Saxon Dictionary* [2], which contains extensive information on the distribution of senses and a considerable amount of sense nesting, and the contemporary Dictionary of Old English (DOE)<sup>3</sup> generally considered the most authoritative dictionary of the language. This paper focuses on the incorporation of a pre-existing dataset describing emotion terms in OE into OldEWN. More precisely, this dataset categorises entries in the OE lexicon of emotions on the basis of the different figurative semantic shifts, if any, which occurred historically between different senses of the same word (this is part of a wider research program in Cognitive Linguistics described in). One of the outcomes of this work of incorporating this dataset into our wordnet has been the development of an extended version of the standard Global Wordnet Association Schema for wordnets<sup>4</sup> as adapted to the exigencies of studying historic languages from a diachronic perspective within a cognitive linguistics inspired framework<sup>5</sup>. Another outcome of the work is to demonstrate the usefulness of the WN organisational approach and schema (with appropriate extensions such as ours) in making datasets such as these in historical lexical semantics more interoperable and potentially easier to disseminate on the basis of principles that are fairly intuitive and easily grasped (notion of synsets with relations like hyponymy between them). In what follows we will show this with the example of OE terms related to the notion of guilt. First, we will describe our workflow, remarking upon the potential of this task for didactic ends. Then we will show some of the OE synsets which we derived for the semantic field of GUILT.

<sup>1</sup> By Old English here we refer to a language spoken in the British Isles from around the fifth century AD up until around the 12th century. It is the earliest ancestor of the English language for which we have written evidence.

<sup>2</sup> We foresee an initial phase of automatic assignment of synsets to word senses, followed by a subsequent phase of correction by Old English experts via a specially designed platform.

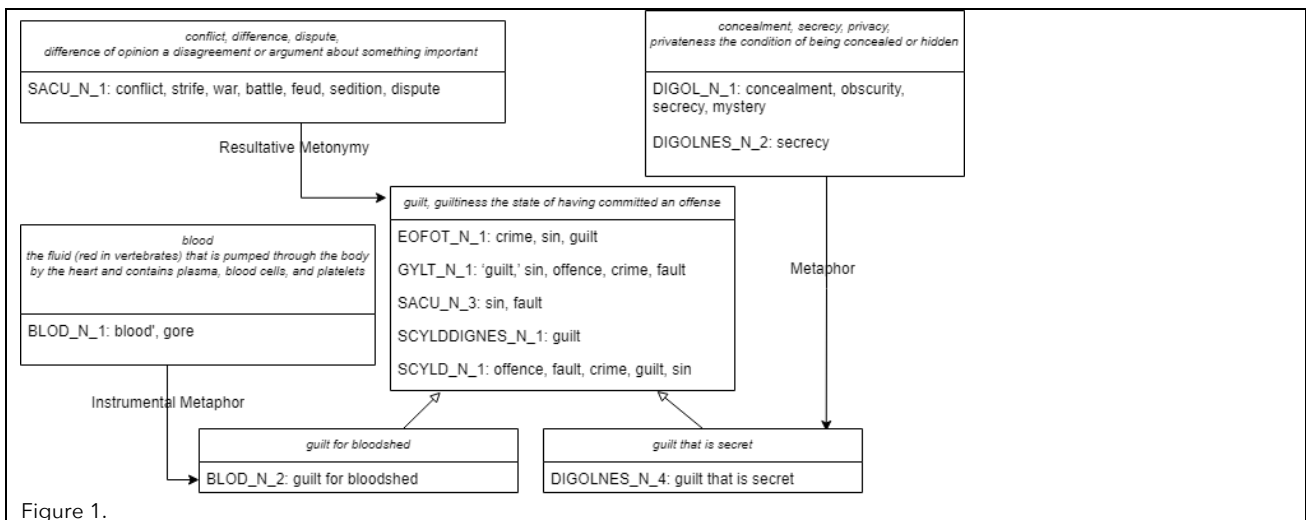
<sup>3</sup> <https://doe.artsci.utoronto.ca/>

<sup>4</sup> <https://globalwordnet.github.io/schemas/>

<sup>5</sup> <https://github.com/anasfkhan81/OldEnglish/blob/main/WN-IELMF-0.DTD>

## 2. WORKFLOW AND EXAMPLE

The task of converting the original Diaz-Vera (DV) dataset with its description of the OE emotion lexicon into a wordnet style lexicon was part of a collaboration between three institutions<sup>6</sup> and involved interns from the University of Siena working under the supervision of experts from CNR-ILC and UCLM. In addition to contributing to the creation of the OldEWN<sup>7</sup>, this task also serves the didactic purpose of introducing the interns both to wordnets and to the Old English lexicon (and more broadly to language resources for ancient languages) as well as to working with lexicographic resources. The task begins with the alignment of the lemmas and descriptions in the original DV dataset (this dataset is described in more detail in [3]) with those in the Clark Hall (CH) dictionary<sup>8</sup>, followed by the enrichment of the CH definitions from the Bosworth Toller dictionary when this is required by the data; in some cases senses are added based on other scholarly sources. Since this part of the wordnet is based on a specialist dataset we include some very rare senses in our synsets; these specialised senses will however be marked as such in the final version of the resource when it is made available. The OE word definitions are subsequently compared with the synset definitions in the Open English wordnet<sup>9</sup> (OEWN), an open source version of the Princeton WordNet. Each definition is matched to its closest synset in case that one exists. Where there exists no such match, the annotator must look for an OEWN hypernym (this usually tends to be quite straightforward). Although this task is currently fully manual, it has given us significant insight into how the CH can best be leveraged for building OE synsets automatically and the many challenges such a task faces. We end this submission with a diagram representing part of the semantic field of GUILT in OE on the basis of the DV dataset structured as a wordnet. Here we show the more ‘core’ terms in the OE guilt vocabulary, in this case nouns such as *eofot*, *gylt*, *sacu*, and *scyld* in their various different senses. Note the extra relationships which we have added to the basic repository of wordnet relations proposed by the GWA. These allow us to signal diachronic sense shift relationships between the senses in synsets and to categorise the kind of figurative (metaphoric, metonymic) shift involved (e.g., from secrecy to a secret guilt).



In our poster we will present a map of the entire semantic field taking into consideration different parts of speech.

## REFERENCES

- [1] Clark-Hall, J. R. A concise Anglo-Saxon dictionary: for the use of students. Swan Sonnenschein & Company, 1894.  
 [2] Bosworth, J., & Toller, T. N. An Anglo-Saxon Dictionary, Oxford, 1889.  
 [3] Khan, A. F., Gómez, F. J. M., González, R. C., Diakoff, H., Diaz-Vera, J. E., McCrae, J. P., O’Loughlin C., Short W.M., & Stolk. S. Towards the Construction of a WordNet for Old English. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 3934–3941, Marseille, France. European Language Resources Association, 2022.  
 [4] Khan, A. F., McCrae, J. P., Gómez, F. J. M., González, R. C., & Diaz-Vera, J. E. Some Considerations in the Construction of a Historical Language WordNet. Global WordNet Conference, 2023.

<sup>6</sup> The Istituto di Linguistica Computazionale (CNR-ILC), the University of Siena, and Universidad de Castilla-la Mancha (UCLM)

<sup>7</sup> The part of the wordnet dealing with emotions will eventually be incorporated into the rest of the OldEWN.

<sup>8</sup> The CH lemmas are already the basis of the lemma list of the contemporary and authoritative Dictionary of Old English.

<sup>9</sup> <https://en-word.net/>

# A digital diplomatic edition of the 10th-century charters of Lucca for Latin corpus linguistics

Timo Korhakangas<sup>1</sup>

<sup>1</sup> Academy of Finland / University of Helsinki, Finland – timo.korkiakangas@helsinki.fi

## ABSTRACT

This paper presents the sampling and edition principles that underlie a digital diplomatic edition of 10<sup>th</sup>-century Latin charters, which was mainly created for corpus-linguistic use. The paper also presents the on-going process of the edition's conversion into a language resource with lemmatization and morphological and syntactic parses and discusses theoretical and methodological issues involved in diplomatically enriched digital corpora of historical texts.

## KEYWORDS

Latin; charters; digital edition; digital diplomatics; corpus linguistics

## 1. INTRODUCTION

This paper presents a recent digital diplomatic edition of 72 10<sup>th</sup>-century charters published as TEI XML in the Archivio della Latinità Italiana del Medioevo (ALIM)<sup>1</sup> in 2021 and describes the on-going process of the edition's conversion into a full-fledged language resource for Latin linguistics. The conversion involves integrating the diplomatic and textual information present in the digital edition into a lemmatized and morphologically and syntactically parsed corpus (treebank). Such an integration entails theoretical and methodological choices that must be justified to enable diplomatically and philologically context-sensitive corpus-linguistic research designs.

The charters of the diplomatic edition come from the *fondo Diplomatico* of the Archivio storico diocesano of Lucca. Thousands of early medieval charters, i.e., private documents written by notaries on single parchment sheets, survive as originals in European, mostly Italian, archives. They are not only a central source for Latin linguistics but also for understanding the social, economic, and administrative history of the Middle Ages. The Archivio storico diocesano of Lucca preserves a patrimony of more than 1,800 documents from before AD 1000 [14]. While the Luccan charters prior to the 10<sup>th</sup> century have been published in modern high-quality editions, the ca. 700 10<sup>th</sup>-century charters were previously only available as incomplete 19<sup>th</sup>-century transcriptions made by Domenico Barsocchini, which contain a large number of errors [3]. Barsocchini also heavily abbreviated and normalized the language of the charters, and he often arbitrarily expanded the original abbreviations without any notice in the text. Due to the considerable importance of charters for various fields of research, the author proposes to provide, along with a diplomatically enriched treebank for linguists, an integrated digital edition in which the treebank annotations are linked to the tokens in order to serve diplomatic and historiographical studies.

The present edition has come about as part of the author's projects on the Latin of early medieval Italian documents, which have utilized corpus-linguistic and digital-philological methods. The edition is the first part of a larger comprehensive edition of the 10<sup>th</sup>-century charters of Lucca that the author plans to complete during his on-going and future projects; an extension of 50 charters is under preparation as of 2023. The author's aim is to utilize the resulting diplomatically enriched treebank as a complement to the Late Latin Charter Treebank (LLCT), a substantial corpus of early medieval charter data (1,040 charters) from the 8<sup>th</sup> and 9<sup>th</sup> centuries, the first two parts of which the author had built between 2011 and 2018 [7]. The second part of the LLCT is also available as a Universal Dependencies version [8].

Medieval charters are not only important for Latin linguistics because of their lexicon, which as such is essential for historical onomastics, toponomastics, and dialectology, but also because of their morphology and syntax. Since charters are formulaic by definition, the linguistic panorama of a charter corpus largely reflects those forms and constructions that happen to be present in the formulae, while the non-formulaic, so-called free parts, which could theoretically display a wider range of linguistic features, are short in early medieval Italian charters [12]. In this respect, charters closely resemble elicited texts, i.e., texts that are elicited from (modern) informants to shed light on the contemporary use of specific linguistic forms and constructions [13]. The linguistic panorama of such texts is equally limited, and the distributions of the linguistic features present in elicited corpora are skewed, but such texts are nevertheless precious to the analysis of the

---

<sup>1</sup> [http://en.alim.unisi.it/dl/fonti\\_documentarie](http://en.alim.unisi.it/dl/fonti_documentarie)

features that they have been elicited for. In the same way, charter corpora can be used to answer carefully planned linguistic research questions. Early medieval charters are all the more important for Latin linguistics because few other original texts survive from the period. With charters, one can be sure that their linguistic appearance derives from the original scribe's pen, not from the mistakes and emendations of an unknown copyist, as is the case with literary texts transmitted through manuscript traditions.

## 2. SAMPLE

The present edition consists of a sample of 72 charters (48,808 tokens) selected evenly from throughout the 10<sup>th</sup> century (Table 1). Five criteria were followed when selecting the charters: i) the charters cover the whole of the 10<sup>th</sup> century in as balanced a way as possible (the *fondo* does not form a continuous time series); ii) only dated original charters and a few copies written relatively soon after their originals are included; iii) only charters in decent physical condition are included, given that extensive lacunae notably reduce the charters' usability for syntactic analysis; iv) the charters represent various types of documents (e.g., sales contracts, leases, exchanges, and donations, the *libellus* being, though, by far the most frequent category); and v) as many different scribes as possible are included. The author transcribed the charters from the original parchments in the archive, but he also utilized digital reproductions that are available upon request from the archive. The compilation of the sample is described in more detail in the preface of *Carte del secolo X dell'Archivio storico diocesano di Lucca* [5].

The purpose of this kind of sampling is to prepare a dataset to be used preliminarily to compare (the Latin of) the charters of the entire 10<sup>th</sup> century with those of the previous centuries, which are accessible through LLCT, before all the 10<sup>th</sup>-century charters are edited. This latter task is likely to take years, and the author wants to perform diachronic comparisons already in his current project. The main reason the author decided to make the sample openly accessible for the research community already at the present stage is that it provides various improved readings compared to those of the 19<sup>th</sup>-century transcriptions.<sup>2</sup>

No	Date	Type	Scribe	No	Date	Type	Scribe
1047	901	libellus	Teudimundus not.	1312	945	exchange	Arnifridi not.
1058	902	trial record	Cospertus not.	1318	947	exchange	Teutpaldus not. iudex dn. regis
1064	903	libellus	Petrus not.	1324	948	libellus	Johannes not. iudex dn. reg.
1082	904	ordination	Petrus not. schab.	1329	949	sale	Bonitio not. et iudex dn. reg.
1092	906	libellus	Floripertus cler. not. schab.	1337	951	exchange	Petrus not. dn. reg.
1095	906	libellus	Andreas not.	1347	952	sale	Arnifridi not. dn. reg.
1099	907	libellus	Petrus not. schab.	1353	953	exchange	Amalbertus not.
1114	908	lease	Andreas not.	1359	954	exchange	Lambertus not.
1120	909	libellus	Richimundus not.	1363	955	libellus	Johannes not. iudex dn. reg.
1124	910	libellus	Richimundus not.	1375	956	libellus	Andreas not.
1134	911	ordination	Fraimundus not.	1383	958	libellus	Fulbertus not. dn. reg.
1140	911	libellus	Anbertus not.	1384	959	exchange	Petrus not. dn. imp.
1146	913	libellus	Fraimundus not.	1386	960	ordination	Leo not.
1148	913	repromission	Petrus not. schab.	1390	961	exchange	Petrus not. dn. reg.
1156	914	evacuation	Ostripertus not.	1394	963	exchange	Petrus not. dn. imp.
1158	915	ordination	anonymous	1395	963	libellus	Petrus not. dn. imp.
1172	916	libellus	Richimundus not.	1398	964	libellus	Petrus not. dn. imp.
1179	917	libellus	Fraimundus not.	1405	968	ordination	Petrus not. iudex dn. imp.
1181	918	ordination	Johannes diac. card. et cantor	1408	968	libellus	Gherardus not. dn. imp.
1186	919	donation	Dominicus not.	1420	970	libellus	Petrus not. dn. imp.
1191	921	ordination	Petrus not. schab.	1427	971	libellus	Raineri not. dn. imp.
1195	923	libellus	Petrus not. schab.	1445	973	libellus	Petrus not. dn. imp.
1199	924	exchange	Teutpaldus not.	1454	974	libellus	Gottfredus not. dn. imp.
1207	927	libellus	Petrus not. schab.	1466	975	libellus	Andreas not. iudex dn. imp.
1213	927	libellus	Johannes not.	1471	976	libellus	Ghisalprandus not. iudex dn. imp.
1225	930	libellus	Ildipertus not.	1497	979	libellus	Petrus not. dn. imp.
1229	933	sale	Winibertus not.	1527	983	libellus	Petrus iudex dn. imp.
1236	936	lease	Petrus not.	1589	984	donation	Johannes not. iudex dn. imp.
1241	936	libellus	Johannes not. iudex dn. regi	1619	987	sale	Ildighisi not. dn. imp.
1246	937	libellus	Roppaldus not.	1633	988	libellus	Gottfredus not. dn. imp.

<sup>2</sup> In addition to the digital TEI XML edition, which is intended for corpus linguists and which is accessible through the ALIM user interface, where it is also visualized according to the ALIM layout ([http://en.alim.unisi.it/dl/fonti\\_documentarie](http://en.alim.unisi.it/dl/fonti_documentarie)) [2], the edition of the sample is also made openly available as a traditional human-readable diplomatic edition in PDF for the benefit of diplomatists and historians [5].



1253	938	exchange	Johannes not. iudex dn. regi	1660	991	libellus	Raineri not. dn. imp.
1262	939	libellus	Amalbertus not. dn. regis	1689	993	libellus	Benedictus not. dn. imp.
1272	940	libellus	Johannes not. iudex dn. regi	1703	995	libellus	Witternus not. dn. imp.
1284	941	sale	Arnifridi not.	1715	997	libellus	Rodilandus not. dn. imp.
1303	943	libellus	Hildegisus not. (sacri palacii)	1740	999	libellus	Rodolfus not. dn. imp.
1308	944	libellus	Petrus not. dn. reg.	1752	1000	libellus	Teutpertus not. dn. imp.

Table 1. The composition of the edition sample. The identifier numbers reproduce those of Barsocchini's edition.

### 3. PRINCIPLES OF THE DIPLOMATIC EDITION

The diplomatic edition predominantly follows the good practices established by the *Chartae Latinae Antiquiores* series, which in turn rely on the Leiden conventions, which were originally designed for papyri but are easily adaptable to charters [6]. The edition of each charter text indicates the abbreviations, the lacunae, the letters that are difficult to read without the consideration of context, as well as the editorial expansions of abbreviations and the restorations of the lacunae when applicable. Original scribal mistakes, corrections, and unexpected linguistic forms that may confuse the modern user are also indicated.

The diplomatic edition presents the text in a form that respects, as closely as possible, the form in which the scribe wished to publish the charter. The only unindicated editorial change concerns the spaces between words: it was considered necessary for NLP purposes to split the words written without breaks into units separated by spaces according to modern conventions (e.g., *in loco* > *in loco* “in (the) place”). Conversely, the original punctuation and capitalization are retained. All editorial interventions except for the spaces are unambiguously indicated using the rules established by the Leiden conventions. For a detailed account of the principles of the edition, see the preface of *Carte del secolo X dell'Archivio storico diocesano di Lucca* [5].

Each charter text is accompanied by contextual metadata concerning: i) its physical description, including the dimensions and quality of the parchment, the identification of material damages and anomalies, as well as the characterization of the scribal hand; ii) the dating, which is based on regnal years and indiction or, in their absence, an estimation based on a comparison with dated charters of persons involved and/or the qualities of the handwriting; iii) the scribe, whose identification typically only poses problems when several scribes have the same name; iv) the factual contents, presented as a short regest; v) *nota dorsalia*, which are the archival notes on the back side of the charter; and vi) potential special characteristics of the charter and its connections with other charters if known.

### 4. PRINCIPLES OF THE DIGITAL EDITION

The diplomatic edition described above was subsequently converted into TEI XML format using regular expressions to transform the Leiden markup into corresponding XML elements and attributes. The resulting digital edition, which adheres to the subset of the TEI elements highlighted by the ALIM Guidelines (v. 1.1, level 3)<sup>3</sup>, was then published in the *Fonti documentarie* section of the *Archivio della Latinità Italiana del Medioevo*. The tagset is also largely compatible with the recommendations of the Charters Encoding Initiative (CEI)<sup>4</sup>, although it does not employ all its possibilities: since normalization was kept to a minimum, only a few items of the TEI element repertoire available for the indication of editorial correction or conjecture were applied [11];[15]. This section briefly discusses the use of the elements that encode the core diplomatic information in the present edition.

The most common editorial intervention is the expansion of abbreviations. The expanded letters are marked with the <ex> element, like with the abbreviation *prbo* for *presbitero* “priest”: pr<ex>es</ex>b<ex>iter</ex>o. The many letters that are discernable only with the help of context are encoded with <unclear>, e.g., iustit<unclear>ia</unclear><unclear><ex>m</ex></unclear><lb/> (Leiden: *iustitiā(m)*) in charter no. 1207 (AD 927), where the last two letters *ia* as well as the superimposed abbreviation mark of the final *-m* can only be deduced from context: the word is expected to be an accusative form of *iustitia* “justice”, and there is only space for two letters before the line break indicated by <lb/> (“line beginning”).

The restorations of lacunae or damaged characters are encoded with <supplied>, e.g., episc<supplied reason="lost">[upa]</supplied>tui, where the entirely damaged letters *upa* can be restored based on other attestations of the genitive form *episcopatus* “bishopric” in the same document. Wider lacunae that cannot be restored are marked with estimates of the number of missing letters, e.g., <supplied reason="lost">[...5–10...]</supplied>. Since the ALIM interface does not currently visualize some Leiden-style diplomatic formatting, the digital edition makes a compromise on the

<sup>3</sup> [https://docs.google.com/document/d/1Xcb8AnmL51jot\\_i-POQ7zjRhHZsDgYpQxNCweR5A4bY/edit#](https://docs.google.com/document/d/1Xcb8AnmL51jot_i-POQ7zjRhHZsDgYpQxNCweR5A4bY/edit#)

<sup>4</sup> <http://www.cei.lmu.de/taglib>



representation of restorations: to render restorations so that they are displayed correctly on the interface, the square brackets of the diplomatic edition have been preserved in the digital edition for the time being, such as [upa] and [...5–10...] above. As for scribal mistakes, <supplied reason="omitted"> is used to indicate haplographies, such as rodolfi <supplied reason="omitted">fi</supplied>lio (Leiden: *rodolfi* <fi>lio), where the scribe has inadvertently missed the repeated syllable *fi*. The element <surplus> encodes dittographies, such as constitui<surplus>ui</surplus>t (Leiden: *constitui*{ui}t), where the scribe has erroneously duplicated the syllable *ui*. Other types of scribal mistakes and corrections as well as unexpected linguistic forms are reported by using critical notes, <note type="critical">, appended to the word in question.

The charters of the present edition display three non-alphabetical signs to convey specific diplomatic information. They are encoded with <g> (“glyph”). The combination <g type="signum">S</g> stands for the *signum manus* of the scribe at the opening of the charter and/or ahead of his subscription (where CEI would utilize the tag <notariusSign>). The combination <g type="crux">+</g> stands for the *signum crucis*, the sign of the holy cross, which the illiterate drew instead of a signature and which clerics used instead of the *S* sign when witnessing a legal act. The symbol “2”, as in <g type="finale">2</g>, is used to encode the curly closing symbol of certain witness and scribal subscriptions. The metadata described in the previous section are presented in each charter’s TEI header using the normal modules in conformity with the ALIM Guidelines.

## 5. TOWARDS MORE CONTEXT-SENSITIVE CORPUS LINGUISTICS

Main-stream corpus linguistics largely ignores the textual-scriptural characteristics of its research data. This may not be a problem with carefully balanced modern-language corpora. The collector of a modern-language corpus can often choose unproblematic texts from among thousands or millions of candidates, whereas the compiler of a historical corpus usually has to do with what is extant. Consequently, the textual peculiarities of the historical texts included in a corpus may essentially bias linguistic conclusions if not taken into due account. Suffice it to mention the expansions of abbreviations and the restorations of lacunae, which derive from the editor’s educated guesses, or the differing readings of manuscript witnesses in the case of transmitted texts. Therefore, it is important to develop methods that enable historical corpus linguistics to cope with complicated textual features. With a charter corpus, this means integrating diplomatic information into linguistic annotation. The following example (see the annotation snippet below) illustrates the solution adopted by the author’s current project.

Prior to linguistic parsing, the text of the diplomatic edition is cleared of diplomatic markup, tokenized, and converted into the CoNLL format so that there is one token per line with sentences separated by a blank line. In CoNLL, each numbered token is followed by eight tab-separated fields that present the linguistic analysis of the token [4]. The author’s project utilizes five fields to indicate lemma, PoS, morphological tag, dependency head, and dependency relation, while the last field (the so-called MISC field) is reserved for the diplomatic information that is transferred from the TEI XML file for each token. This field reproduces the token with its XML tags just as it appears in the digital edition.

This diplomatically enriched CoNLL representation of the text is fed into a linguistic parser, here UDPipe2, which has been trained on a language model based on the Universal Dependencies release of LLCT2 (latin-llct-ud-2.10-220711)<sup>5</sup>, which appends a set of linguistic analyses to each token. The resulting CoNLL-U parse can be then queried to answer linguistic research questions, paying simultaneously attention to the relevant diplomatic and textual features. For example, the diplomatic annotation pr<ex>es</ex>b<ex>ite</ex>r appended to the token *presbiter* permits the delimitation of queries to the actual text of the charter, i.e., *prbr*, which excludes the possibility to use the token as evidence of the wide-spread vocalism *i* instead of *y*, even though this is the most common spelling when the word appears unabbreviated. The word appears in the CoNLL-U file in its expanded form *presbiter* in order to ease NLP tasks (parsing and querying).

The snippet below displays the CoNLL-U parse of a subscription clause that corresponds to the following Leiden formatting: [+]  
ego cristianu pr(es)b(ite)r in hanc\* ordinatione consensi et subs(cripsi) “[+] I, Christian, the priest, agree on this ordination and subscribe to it” (charter no. 1134, AD 911). The asterisk replaces a footnote that conveys the information that the scribe has subsequently added the *h* between the lines (see line 6 below).

```
1 + + PUNCT _ _ 8 punct _ <supplied reason="lost">[<g type="crux">+</g>]</supplied>
2 ego ego PRON _ Case=Nom|Gender=Masc|Number=Sing 3 det _ <unclear>ego</unclear>
3 cristianu Christianus NOUN _ Case=Acc|Gender=Masc|Number=Sing 8 nsubj _ <unclear>cr</unclear>istianu
4 presbiter presbyter NOUN _ Case=Nom|Gender=Masc|Number=Sing 3 appos _ pr<ex>es</ex>b<ex>ite</ex>r
5 in in ADP _ _ 7 case _ in
```

<sup>5</sup> <https://ufal.mff.cuni.cz/udpipe/2/models>

6 hanc hic DET \_ Case=Acc|Gender=Fem|Number=Sing 7 det \_ hanc<note type="critical">h nell'interlinea superiore</note>  
7 ordinatione ordinatio NOUN \_ Case=Acc|Gender=Fem|Number=Sing 8 obl \_ ordinatione  
8 consensi consensio VERB \_ Aspect=Perf|Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin|Voice=Act 0 root \_  
consensi  
9 et et CCONJ \_ \_ 10 cc \_ et  
10 subscripsi subscribo VERB \_ Aspect=Perf|Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin|Voice=Act 8 conj \_  
subs<ex>cripsi</ex><lb/>

The practice delineated above beats the one utilized in the LLCT treebanks, in which only the words whose inflectional endings are abbreviated or restored are tagged, so that they can be excluded from linguistic processing, although they are present as nodes in the syntactic trees [10];[9]. With the solution described here, the researcher can focus precisely on the matter that is not abbreviated or restored. Moreover, the information on scribal mistakes, such as haplographies and dittographies, enables an additional analysis of the writing habits of individual scribes. Another improvement in respect to LLCT is the possibility to detect the token's position within the line of the original parchment thanks to the <lb/> tag. The approaching line ending sometimes challenged the scribe who wished not to divide a word between two lines, hence the unusual abbreviations or even mistakes at line breaks.

The linguistic annotation in the CoNLL-U format predominantly meets the needs of corpus linguists. However, even diplomatists and historiographers may benefit from linguistic information, especially lemmatization, which allows them to identify all the occurrences of historically relevant key terms with full recall in spite of the huge spelling and inflectional variation of charters. Therefore, at a later stage of the project, the linguistic annotations will be linked word for word to the diplomatic edition by using, for example, the CTS URN scheme that has been successfully exploited by the Perseus Digital Library to combine the words of the works of classical authors to respective dictionary entries or other resources [1]. Thus, the author proposes a linkage between each CoNLL-U line and TEI XML token, which will be aligned through the last field of the CoNLL-U line. In this way, the lemma, morphological analysis, and the syntactic dependency relation of each word can be consulted, preferably through the ALIM interface, provided that its upcoming versions will support such a linkage.

## 6. ACKNOWLEDGEMENTS

The author wishes to thank Prof. Francesco Stella and the whole ALIM team for their invaluable help in preparing the final version of the digital edition. This paper is based on research funded by the Foundation *Institutum Romanum Finlandiae* and the Academy of Finland (grant no. 315176).

## REFERENCES

- [1] Almas, Bridget, Alison Babeu, and Anna Krohn. «Linked Data in the Perseus Digital Library». ISAW (Institute for the Study of the Ancient World) Papers 7.3, 2014. <http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/almas-babeu-krohn/>
- [2] Archivio della Latinità Italiana del Medioevo, «Carte del fondo Diplomatico dell'Archivio storico diocesano di Lucca», ed. Timo Korhakangas, 2021. [http://en.alim.unisi.it/dl/fonti\\_documentarie](http://en.alim.unisi.it/dl/fonti_documentarie)
- [3] Barsocchini, Domenico. *Memorie e documenti per servire all'istoria del Ducato di Lucca* 5:3, Bertini, 1841.
- [4] Buchholz, Sabine, and Erwin Marsi. «CoNLL-X Shared Task on Multilingual Dependency Parsing». Márquez, Lluís, and Dan Klein, eds., Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). Association for Computational Linguistics, 2006: 149–164. <https://aclanthology.org/W06-2920.pdf>
- [5] Carte del secolo X dell'Archivio storico diocesano di Lucca, ed. Timo Korhakangas, 2021. <http://doi.org/10.5281/zenodo.4965245>
- [6] *Chartae Latinae Antiquiores*, Facsimile-edition of the Latin Charters, 2nd Series: Ninth Century, voll. 50–118, ed. Guglielmo Cavallo et al. Urs Graf Verlag, 1997–2019.
- [7] Korhakangas, Timo. «Late Latin Charter Treebank: contents and annotation». *Corpora* 16 (2021): 191–203.
- [8] Korhakangas, Timo, Flavio Massimiliano Cecchini, and Marco Passarotti. «Late Latin Charter Treebank». Dan Zeman et al., eds., *Universal Dependencies 2.6*, Institute of Formal and Applied Linguistics (ÚFAL), 2020. [https://github.com/UniversalDependencies/UD\\_Latin-LLCT](https://github.com/UniversalDependencies/UD_Latin-LLCT)
- [9] Korhakangas, Timo, and Marco Passarotti. «Challenges in annotating Medieval Latin charters». *Journal of Language Technology and Computational Linguistics* 26 (2011): 103–114. <https://jllc.org/content/2-allissues/13-Heft2-2011/16.pdf>
- [10] Korhakangas, Timo, and Matti Lassila. «Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material». Mambrini, Francesco, Marco Passarotti, and Caroline Sporleder, eds., Proceedings of The Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3). Bulgarian Academy of Sciences, 2013: 61–72. <http://bultreebank.org/wp-content/uploads/2017/06/ACRH-3Proceeding.pdf>

- [11] «Representation of Primary Sources». TEI P5: Guidelines for Electronic Text Encoding and Interchange, version 4.5.0., Text Encoding Initiative Consortium, 2022. <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html>
- [12] Sabatini, Francesco. «Esigenze di realismo e dislocazione morfologica in testi preromanzi». *Rivista di cultura classica e medievale* 7 (1965): 972–998.
- [13] Senft, Gunter. «Elicitation». Senft, Gunter, Jan-Ola Östman, and Jef Verschueren, eds., *Culture and Language Use*. Amsterdam, 2009: 105–109.
- [14] Todros, Gabriela. «Brevi considerazioni sul patrimonio archivistico dell’Arcidiocesi lucchese». Pagano, Sergio, and Pierantonio Piatti, eds., *Il patrimonio documentario della chiesa di Lucca: prospettive di ricerca*. Galluzzo, 2010: xi–xiv.
- [15] Vogeler, Georg. «Die Text Encoding Initiative (TEI) als Werkzeug des Urkundeneditors – Erfahrungen und Desiderate». Fees, Irmgard, Benedikt Hotz, and Benjamin Schönfeld, eds., *Papsturkundenforschung zwischen internationaler Vernetzung und Digitalisierung: Neue Zugangsweisen zur europäischen Schriftgeschichte*. Göttingen, 2015: 1–11.

# On why and how we should build a catalogue of software products for digital scholarly editing

Chiara Martignano<sup>1</sup>

<sup>1</sup>Università degli Studi di Firenze, Italy – chiara.martignano@unifi.it

## ABSTRACT

The increased number of software products currently available for digital scholarly editing suggests it would be convenient to create a dedicated catalogue, in order to help scholars and institutions choose up-to-date tools and to keep track of the changes that occur in the field. Moreover, a catalogue of software products for digital scholarly editing could prove itself to be a valuable resource to study the technological trends of digital philology from a historical perspective.

In order to build such a catalogue many aspects need to be carefully considered. In this paper, we reflect on the potential scientific value of the catalogue and present the issues that have to be studied, using TAPoR [9] as an example of tools collection, from which one could draw inspiration.

## KEYWORDS

Catalogue; digital scholarly editing; digital editing tools; digital editing software

## 1. WHY

In the field of digital philology a large number of tools has been developed through the years and keeps being developed in order to assist (or substitute) editors in the making of a digital scholarly edition. At the moment there is no catalogue available that focuses on the field of digital philology and keeps track of all the software products that exist or have existed so far in this particular field of the digital humanities.

The two different catalogues of scholarly digital editions, created respectively by Patrick Sahle [8] and Greta Franzini [2], have proven to be indispensable research tools that allow scholars to keep an eye on the field of digital philology from a wider perspective. Both catalogues register the information about the tools and technologies used within the digital editions, when it is available.<sup>1</sup> However, the variety of tools available and the increased interest in this branch of the digital humanities, calls for a new dedicated catalogue.

A catalogue of software products and digital tools for scholarly editing would be very useful for various reasons. First of all, it would help potential users, both libraries, archives and other cultural institutions, as well as individual scholars, find the tools that suit their research goals the best. Second, it would facilitate the community of scholars to keep track of the changes and developments of the different tools. Third, the catalogue would be a very interesting research tool to analyse the field of digital philology and its history from a more technical point of view.

As many of us know, software products and tools developed in this field tend to stay around for a few years and then silently disappear. This phenomenon is very frequent both because of the inevitable and rapid changes that occur in technology and because very often their development and maintenance rely on project funds that are time-limited. In other cases, it is difficult to infer the status of a software, whether it is still maintained or if it is compliant with up-to-date technologies, from its documentation published online or from articles. Another common case is when a software has been rebooted by a different development team or agency and even under a new name. All these problems make the evaluation and selection of tools really difficult, especially for young scholars who are new to the field and don't know the "stories" behind the different projects that developed the tools.

Having a catalogue that focuses on the software products of the research in the field of digital philology, would make it possible to study the technological trends that are now dominating and that have dominated in the past years the practice of scholarly editing. One could use the data derived from the catalogue to identify the software products that were more frequently adopted and to trace the most common editorial workflows (or "recipes") during a given period of time. For example: between 2010 and 2015 most digital scholarly editions were produced using the authoring software A, then analysed with software B and published as web pages with software C. It would be possible to conduct a sort of "archeological" research, keeping an historical record on all the software products that have been created and their impact on the field.

---

<sup>1</sup> In Franzini's catalogue, for example, the information about the tools and technologies used is available for 178 out of 320 records.

## 2. THE EXAMPLE OF TAPOR

In this section we present TAPoR (Text Analysis Portal for Research) [9], a collection of research tools for digital humanists, as an example to reflect on which scientific requirements and features a catalogue of tools for digital scholarly editing should have. TAPoR, developed by Geoffrey Rockwell and Milena Radzikowska with the support of the Arts Resource Centre at the University of Alberta, is available as a database and online platform. Originally focused on text analysis tools, as the name suggests, the collection was later expanded to include the DiRT<sup>2</sup> Directory [5], and other tools that work on non-textual data and provide services used by digital humanists, such as publishing tools, GIS tools and communication tools.<sup>3</sup>

The collection currently includes 1645 records, categorized according to the TaDiRAH<sup>4</sup> taxonomy [1], which comprehends the following main activities: analyzing, capturing, creating, disseminating, enriching, interpreting and storing. In the taxonomy, each of the aforementioned activities comprehends a set of subactivities, which are further divided into additional categories. For example, under the activity “enriching” one can find the subactivity “editing”, and under “editing” the activities “adding”, “correcting”, “merging” etc. The online platform offers an overview of the whole collection in the home page and a detailed description for each record, which includes:

1. general information, i.e. the name of the tool, the date of when the record was last updated, the link to the tool’s official web page and a brief description;
2. information about the creator of the tool;
3. attributes to describe the main properties of the tools, like the type of licence under which the tool is available, and the TaDiRAH categories that best describe the tools functionalities and features;
4. tags and related tools to help the user browse the collection;
5. ratings and user comments.

Figure 1. Webpage dedicated to the tool eLaborate within the TAPoR platform.

The TAPoR platform also offers an advanced research interface that allows users to filter tools according to further properties, for example the “ease of use”, the development status and possible warnings.

<sup>2</sup> DiRT stands for Digital Research Tools. The directory was developed by Digital Humanities at Berkeley.

<sup>3</sup> For a complete list of the types of tools included in TAPoR see the “About” page of the project: [https://tapor.ca/pages/about\\_tapor](https://tapor.ca/pages/about_tapor).

<sup>4</sup> TaDiRAH stands for Taxonomy of Digital Research Activities in the Humanities.

Filters: Narrow the Selection of Tools

Type of analysis	<input type="text"/>	Type of license	<input type="text"/>	Background Processing	<input type="text"/>
Web Usable	<input type="text"/>	Ease of Use	<input type="text"/>	Warning	<input type="text"/>
Usage	<input type="text"/>	Tool Family	<input type="text"/>	Historic Tool (developed before 2005)	<input type="text"/>
Compute Canada	<input type="text"/>	Link to Recipe	<input type="text"/>	TaDiRAH Goals	<input type="text"/>
TaDiRAH Methods	<input type="text"/>	Historic Tool	<input type="text"/>	Tag	<input type="text"/>
Type	<input type="text"/>				

Figure 2. The research form in the TAPoR platform.

### 3. HOW

In order to create a catalogue of software products for digital scholarly editing, many aspects need to be carefully evaluated. First of all, from a scientific point of view it is necessary to define which kinds of software products can be part of the catalogue. Should the catalogue focus on tools that are specifically designed for digital scholarly editing or should it include general purpose tools, like Wordpress, that are frequently used in the field of digital philology? Second, it is indispensable to identify a proper set of categories to describe the tools goals. The TaDiRAH taxonomy can be a viable option. However, potential users may need a set of categories that better reflect the tasks they usually undertake to create a scholarly edition, for example “transcription of primary sources” and “collation”. Third, it is necessary to establish which properties for the description of a software product are most relevant to scholars in the field of digital scholarly editing. Information about the creator, the licence, the programming languages and technologies used as well as the creation date and (potentially) when the software development and/or maintenance have been interrupted must be included in the catalogue. Other “technical” properties can be the same used in the TAPoR collection to describe the usage of the software (for example: “Web usage” and “Background processing”). Additionally, it is indispensable to record the formats of the input and output data, for this information can help recognise possible workflows. In order to share examples of how the tools can be applied, the catalogue should preserve for each record the links to the digital scholarly editions that have used it. Further properties can be identified through surveys, following the example of Franzini et al. [4], in order to detect the attributes that the scientific community considers most useful. Finally, it would be very helpful to register users’ comments, to crowdsource up-to-date information about the status of the different tools.

### 4. CONCLUSIONS

A catalogue of software products for digital scholarly editing could be a valid instrument and an effective research tool. Above all, if the scientific community decided to invest time and resources in the realization of such a catalogue, it would affirm the importance of building software as a vital scholarly activity in the field of digital humanities. Many of the tools that are available for scholarly editing purposes are built by scholars “to make the general task of scholarship easier for other scholars” [6]. A dedicated catalogue would add value to the work done by these scholars “who have turned to building, hacking and coding as part of their normal research activity” [*ibid.*].

Many factors concerning the development and the maintenance of the catalogue still need to be considered also from a practical point of view. The catalogue could be developed as a “curated list”<sup>5</sup> in the TAPoR collection or as an independent platform. In that case, one should define a plan to make the cataloguing activity sustainable and durable and determine the technologies to build the database and its online website, considering that it would be very convenient for potential users to have access to an API as well. Hopefully, this paper will help start the conversation about the convenience of the catalogue and how it should be built.

<sup>5</sup> TAPoR offers lists of tools that focus on a given area or theme of the digital humanities and that are curated by associated editors.

## REFERENCES

- [1] 'ACDH\_CH\_Vocabularies: TaDiRAH: Taxonomy of Digital Research Activities in the Humanities'. n.d. Accessed 20 February 2023. <https://vocabs.dariah.eu/tadirah/en/>
- [2] 'Dig-Ed-Cat'. n.d. Accessed 9 May 2022. <https://dig-ed-cat.acdh.oeaw.ac.at/>
- [3] Franzini, G., S. Mahony, and M. Terras. 2016. 'A Catalogue of Digital Editions'. In: Pierazzo, E and Driscoll, M, (Eds.) *Digital Scholarly Editing: Theories and Practices*. (Pp. 161-182). Open Book Publishers: Cambridge, UK. (2016), edited by E. Pierazzo and M. Driscoll, 4:161–82. Cambridge, UK: Open Book Publishers. <https://doi.org/10.11647/OBP.0095>
- [4] Franzini, Greta, Melissa Terras, and Simon Mahony. 2019. 'Digital Editions of Text: Surveying User Requirements in the Digital Humanities'. *J. Comput. Cult. Herit.* 12 (1): 1:1-1:23. <https://doi.org/10.1145/3230671>
- [5] Grant, Kaitlyn, Quinn Dombrowski, Kamal Ranaweera, Omar Rodriguez-Arenas, Stéfan Sinclair, and Geoffrey Rockwell. 2020. 'Absorbing DiRT: Tool Directories in the Digital Age'. *Digital Studies / Le Champ Numérique* 10 (1). <https://doi.org/10.16995/dscn.325>
- [6] Ramsay, Stephen, and Geoffrey Rockwell. 2012. 'Developing Things: Notes toward an Epistemology of Building in the Digital Humanities'. *Debates in the Digital Humanities*, 75–84.
- [7] Rockwell, Geoffrey, Stéfan Sinclair, and Milena Radzikowska. 2020. 'Text Analysis Portal for Research (TAPoR) Archive'. ERA. 1 January 2020. <https://doi.org/10.7939/r3-tnha-4924>.
- [8] Sahle, Patrick. n.d. 'A Catalog of Digital Scholarly Editions'. Accessed 17 August 2020. <http://www.digitale-edition.de/>.
- [9] 'TAPoR'. n.d. Accessed 20 October 2022. <https://tapor.ca/home>



# Database Allegorico Dantesco

Matteo Maselli<sup>1</sup>

<sup>1</sup> Università degli Studi di Macerata, Italia – m.maselli2@unimc.it

## ABSTRACT

Il presente lavoro prende in considerazione, illustrandone struttura e funzioni di base, il Database Allegorico Dantesco (DAD), un nuovo *repository* sull'allegorismo della *Divina Commedia*.

## PAROLE CHIAVE

*Divina Commedia*; Dante Alighieri; allegoria; *repository*

## 1. INTRODUZIONE

Il contributo in questione presenterà criticamente il [Database Allegorico Dantesco](#) (DAD), un *repository* digitale in corso di sviluppo dedicato alla raccolta e all'interrogazione di testi accademici inerenti allo studio dell'allegorismo della *Divina Commedia* di Dante Alighieri (Figura 1). Particolare attenzione verrà riservata alla struttura di base del DAD, alle sue funzioni e ai principi metodologici seguiti per la configurazione di tale risorsa.

Il progetto è parte integrante della ricerca dottorale di chi scrive queste pagine – Tutor: Laura Melosi, Giuseppe Ledda – ed ha beneficiato del sostegno economico dell'Università degli Studi di Macerata e dell'University of Notre Dame (il lavoro si inserisce nel The William & Katherine Devers Program in Dante Studies). Sostenitrice del progetto è anche l'University of Oxford che ha messo a disposizione alcuni fondi bibliotecari ritenuti pertinenti all'archiviazione nel database.

Prima di procedere oltre si sente la necessità di puntualizzare che per la realizzazione del progetto qui presentato sono state privilegiate comuni logiche informatiche, da non ritenersi tuttavia banali o superficiali ma da intendersi al contrario come manifestazione di essenzialità e funzionalità. Si vuole cioè dire che la lunga fase di configurazione del database e del suo popolamento con testi preventivamente selezionati, fisicamente recuperati e manualmente trascritti in un formato operabile dalla macchina è stata preceduta da un altrettanto intenso periodo di riflessione al termine del quale si è deciso di privilegiare, come si dirà meglio successivamente, dei principi di base in linea con l'uso quotidiano delle più comuni risorse tecnologiche e dei benefici da queste consentiti.



Figura 1. Logo del Database Allegorico Dantesco.

## 2. PROSPETTO GENERALE

Il DAD è stato pensato e programmato nelle sue funzioni di base per servire sia da strumento di studio per neofiti delle questioni in esso trattate che come mezzo rivolto a studiosi già edotti delle complesse caratteristiche dell'allegorismo del poema dantesco, che potranno così essere ulteriormente approfondite.

Il materiale presente nel database riguarda dei canti della *Divina Commedia* scelti perché particolarmente importanti per il discorso sull'allegorismo e sull'allegoresi del poema: *Inf.* IX, *Inf.* XVII, *Purg.* VIII. Per ognuno dei tre canti sono stati predisposti dei *folder* di raccolta suddivisi in macroaree tematiche: Egesi storica, *Lecturae Dantis*, Saggi & articoli scientifici. Se nelle ultime due trovano posto contributi di epoca contemporanea, nella prima confluiscono lavori pre-novecenteschi con particolare attenzione ai commenti ai canti scritti dal XIV secolo in poi.

Per rendere la risorsa utile anche a studiosi non interessati in prima istanza ai tre canti suelencati, è stata poi predisposta una quarta area tematica – la Sezione monografica – che accoglie circa 300 testi che discutono dell'allegoria dantesca nelle

sue linee generali o in riferimento a caratteristiche specifiche. Tali contributi sono ritenuti dagli specialisti i migliori studi sull'allegoria rilasciati dalla dantistica internazionale nel corso del XX e XXI secolo. La Sezione monografica propone dunque la più completa collezione digitale di testi accademici *open access* sull'allegoria della *Divina Commedia* oggi presente online.

Attualmente – aprile 2023 – sono di libera consultazione quasi 500 testi, per un totale di oltre 8000 pagine trasposte digitalmente e riferite a più di 250 autori diversi, con una percentuale maggiore di voci italiane e inglesi e una più piccola rappresentanza di esponenti francesi e spagnoli.

Con riferimento specifico a quest'ultimo prospetto si precisa che, nel rispetto del *copyright* vigente sui testi pubblicati, laddove la riproduzione digitale eccede il 15% del materiale analogico, si fa riferimento a prodotti già gratuitamente e integralmente disponibili online (*fair use*). Rispetto a quest'ultimi il DAD presenta non indifferenti vantaggi. Oltre a combattere la dispersività dei dati, non esistono aggregatori web che accolgono nello stesso luogo tutto il materiale che sarà catalogato nella risorsa qui discussa. Ciò permetterà di ovviare al grosso limite di imporre delle lunghe ricerche, alle volte complesse o addirittura inattuabili, in più aree web. Inoltre, i testi già fruibili online sono spesso inerti, non consentendo cioè un'interoperabilità come quella che si vuole invece proporre con il DAD.

### 3. I PRINCIPI ALLA BASE DEL DAD

Il DAD risponde a due principi di base, relativi rispettivamente all'ipertestualità tra il materiale archiviato e alla correlazione tra dati.

Il ricorso a questi modelli concettuali ha obbligato accurate operazioni preliminari per l'acquisizione digitale dei testi. In tal senso si è fatto uso di uno scanner OCR (*Optical Character Recognition*) con il quale è stato possibile acquisire i testi selezionati tutelandone l'integrità strutturale. Dopo aver opportunamente corretto ogni testo con programmi di videoscrittura, sono stati integralmente riscritti e marcati in XML rispettando con attenzione le *Guideline* della TEI. Il formato finale presenta non pochi vantaggi, non solo in termini di conservazione – poiché per i testi del DAD si è fatto uso di un formato *standard* sarà possibile usufruire degli stessi su più *device* e trasportare eventualmente i dati archiviati in altre infrastrutture web per contrastare così l'obsolescenza informatica – ma anche e soprattutto nell'ottica della rimodulazione del prodotto di *output*. La struttura nidificata del testo in XML ha permesso infatti l'inserimento di particelle aggiuntive (*tag*) con le quali attribuire caratteristiche ulteriori all'area testuale interessata. Attraverso i *tag* è cioè possibile rendere il testo più dinamico e rispondente a sollecitazioni esterne – come l'inoltro di ricerche mirate.

Nell'estesa casistica dei *tag* si è ritenuto vantaggioso ricorrere soprattutto al marcatore `<ref target="..."> ... </ref>`. La sua inclusione nello schema della trascrizione XML consente l'istituzione di rapporti intertestuali interattivi nella documentazione catalogata ricreando così una consultazione testuale multi-prospettica. Pertanto, l'utente che consulterà le schede del DAD troverà in corrispondenza di citazioni dirette nel corpo del testo o di rimandi bibliografici indiretti in nota dei collegamenti ipertestuali alle opere menzionate che potranno così essere immediatamente esaminate (figura 2).

```
<note place="bottom">Tra gli studi del Singleton non divulgati in versione italiana, ricordiamo
<ref target="https://www.databaseallegoricodantesco.it/view.php?id=446">The irreducible Dove,
in «Comparative Literature», IX (1957), 2, pp. 132-133</ref>, che è risposta a
<ref target="https://www.databaseallegoricodantesco.it/view.php?id=344"> R.H. GREEN,
Dante's «Allegory of poets» and the Medieval Tbeory of Poetic fiction nella stessa rivista,
pp. 118-128)</ref></note>.
```

Figura 2. Esempio di nota a piè di pagina scritta in XML e contenente dei rimandi ipertestuali ad opere archiviate nel DAD.

Il confronto simultaneo, che rimanda al principio dell'ipertestualità prima menzionato, mette a disposizione dei navigatori più voci critiche sull'oggetto ricercato permettendone così una valutazione complessiva che tenga conto anche di posizioni contrarie a quelle criticamente dominati (viene dunque incoraggiato l'insorgere di un esteso *networking* esegetico sull'allegoria dantesca). Affinché possa garantirsi una funzione di questo tipo è stato tuttavia necessario accertarsi che il testo chiamato in causa da un dato autore fosse stato regolarmente acquisito e messo online nel *database*. Ciò ha comportato un'avveduta scelta dei parametri per il reperimento del materiale archiviato.

Nel non facile discernimento dei lavori meritevoli di selezione – difficoltà dovuta tanto all'esteso arco cronologico tenuto in considerazione quanto all'ancor più ampia bibliografia disponibile – si è data priorità all'autorevolezza dello studio e/o del suo proponente e all'alto tasso di ricorsività dei testi tra opere indipendenti dovuto a citazioni dirette e/o indirette. Testi criticamente importanti e storicamente ricorsivi sono infatti particolarmente propensi ad essere soggetti a fenomeni d'intertestualità.

L'ipertestualità riguarda anche gli indici delle schede laddove i testi digitalizzati risultino composti da più unità interne (capitoli, paragrafi, sottoparagrafi). Il sommario interattivo riporta infatti in automatico alle varie ripartizioni con collegamenti ad esse diretti. A prescindere dalla tipologia dei lavori scelti, ogni scheda proporrà poi l'area dei metadati, quella del testo e le note. Nella prima sono riportate le dovute informazioni bibliografiche per un riconoscimento univoco dell'opera in consultazione; la seconda accoglie il corpo del testo corrispondente esattamente all'originario prodotto analogico, mentre le note si troveranno a piè di pagina e saranno raggiungibili con dei puntatori interattivi (figura 3).

## Typology in the Divine Comedy [Alan Clifford Charity]

---

### Table of contents

1. [Figural realism and the state of souls after death](#)
2. [Dante and the aesthetes: the typology of death](#)
3. [Prophecy and the typology of redemption](#)
4. [Conclusion](#)

Indice

Dati bibliografici

Autore: Alan Clifford Charity

Tratto da: Events and their afterlife. The dialectics of christian typology in the Bible and Dante

Metadati

Editore: Cambridge University Press, Cambridge

Anno: 1987

Pagine: 167-261

Testo

The statements from Baumgartel which we have just quoted are applicable to the Commedia too, 'für uns geschrieben, uns zur Warnung, uns zum Trost'. It is the whole of my aim in this final part to show how the Commedia's typology is 'applied' to that purpose of warning and comfort. I hope that by treating the subject here in the context of biblical typology, light will be cast back upon the Bible's use of typology, whose potential is here developed in a direction literally 'extraordinary' without involving fundamental change in its rationale. And I hope too to be able to show that the Comedy gains no less from this concatenation-or rather, that its criticism gains. For in the perspective which the Bible's use of typology gives us, we can go a great way towards overcoming the idea which amounts almost to a fixation in Dante criticism, that the 'allegorical meaning' (and the extent to which this is co-terminous with 'typological meaning' will be discussed later) is a subject for special study, something apart even, in the view of many critics, from the interpretation of the Comedy's 'thought'.

Nota

The Divine Comedy is a poem about conversion. It is, no doubt, incidentally about much else. But that its main subjects: becoming a Christian. The poem being directed (in current parlance, one might say 'geared') to the conversion of the world, or society, or his readers, its author works out his purpose as best he can and as, perhaps, he best can, by narrating his own. It is this enterprise which makes his poem significant in the history and theory of typology. It is this theme, in its working out through typology, which relates the Comedy, suggestively, to the Bible.

Figura 3. Esempio di scheda del DAD con evidenziazione di indice, metadati, testo e note.

Poiché si punta a rendere l'esperienza di studio permessa dal DAD prossima a quella che contraddistingue un ambiente digitale di alto apprendimento si è deciso di implementare anche le modalità di consultazione dei testi. Attraverso un'apposita *toolbar* (figura 4) l'utente potrà gestire dei comandi per annotare qualsiasi testo con commenti personali e sottolineare passaggi da lui ritenuti particolarmente importanti (si è voluto in parte riprodurre la logica delle *Web Annotations* promosse dal W3C). Per salvaguardare l'integrità dei lavori originali, impedendo che le modifiche apportate dal singolo visitatore condizionino la lettura di altri utenti, è stata creata una *workstation* integrata al database. Registrandosi gratuitamente e loggandosi con delle proprie credenziali (indirizzo e-mail + password) si accederà così ad un'area di lavoro personale in cui salvare tutti gli interventi eseguiti sui testi senza che gli stessi compaiano negli originali.

Evidenzia il testo selezionato
Reset Testo Evidenziato
Aggiungi Nota
Salva Nota

Figura 4. Toolbar presente nelle schede del DAD con la quale selezionare varie opzioni di segnatura del testo (evidenziazione e/o annotazione).

Concettualmente più complessi sono gli accorgimenti richiesti per far seguito al principio della correlazione. Con esso si intende il suggerimento in automatico da parte del *software* di voci non considerate dall'utente in fase di ricerca. Si ipotizzi che in una *lectura Dantis* di *Inf. IX*, digitalizzata e inserita nel DAD, un commentatore di formazione latina abbia discusso di Megera, Aletto e Tesifone, i tre mostri che invocano Medusa per sbarrare la strada a Dante, e nel far questo sia ricorso alla sola terminologia della mitologia romana che identificava i tre personaggi come «Furie». A causa di questa scelta lessicale, qualora si andasse a ricercare il termine «Erinni», sinonimo delle Furie ma di discendenza greca, il motore di ricerca non restituirebbe all'utente la *lectura Dantis* in questione. Poiché una simile selezione grava sulla qualità della ricerca condotta dall'utente, che con ogni probabilità ha interesse ad analizzare anche i testi in cui si parla di «Erinni», si sono approntati degli espedienti per ovviare all'estromissione tematica di dati significativi. Nella pagina di *back-end* del *database*, alla quale, per tutelare l'integrità della risorsa, può accedere solo l'amministratore del sito, è stata predisposta una tabella delle ricerche correlate in cui istituire delle relazioni tra termini per risolvere in automatico il problema di omissione nei risultati indicizzati dal *software* (figura 5).

**Gestione delle ricerche correlate**

Parametro di Ricerca

Furie

Termini Correlati

Erinni

Inserire i termini di ricerca correlati separati da ; (punto e virgola)

Aggiungi

Figura 5. Tabella delle ricerche correlate presente nell'area di *back-end* del DAD.

A seguito della compilazione del form con appropriati parametri di ricerca e termini correlati – operazione espandibile senza alcune limitazioni anche a più termini simultanei – il software proporrà dei suggerimenti di voci alternative a quella ricercata cliccando sulle quali si verrà reindirizzati in automatico alla lista completa dei testi del DAD che le contengono. Il motore di ricerca è stato inoltre settato per consentirgli un preciso riconoscimento del valore di una o più keyword inserite nella maschera di query in modo da assicurare un'assoluta pertinenza degli esiti restituiti dalla macchina. Per cui, se ad esempio viene digitato il termine «metafisica» si verrà rinviati alle schede che riportano la sua esatta dicitura tralasciando in automatico quelle in cui si menziona il solo «fisica», le stesse schede che sarebbero state invece proposte qualora si fosse tarata la relazione input-output del motore di ricerca sull'identificazione delle sole lettere dei lemmi («fisica» è infatti parte di «metaFISICA»).

Sempre in merito alle possibilità offerte dalla ricerca, si può selezionare il percorso in cui condurre l'indagine e decidere se espanderla a tutto il corpus del DAD o limitarla a specifiche categorie/sotto-categorie, con la premessa comunque che saranno opportunamente indicati in una sezione pertinente («Archivio») tutti i testi, con relativi luoghi di conserva, caricati nel database. Inoltre, poiché le note a piè di pagina sono per la dantistica un ottimo esempio di «meta-saggi», l'utente potrà circoscrivere la sua ricerca anche solamente a queste voci testuali.

#### 4. IMPLEMENTAZIONI

Si conclude sottolineando la natura *in fieri* del progetto e di come, pertanto, nel prossimo futuro non siano da escludere modifiche contenutistiche e strutturali, anche di grande portata. Si potrebbe, ad esempio, valutare l'implementazione di un'ontologia per la classificazione delle allegorie (operazione che richiederebbe però la collaborazione con informatici umanistici particolarmente esperti). Anche la visualizzazione delle schede potrebbe giovare di accorgimenti mirati come ad esempio la possibilità di un inquadramento cronologico che disponga i testi consultabili secondo veri ordini temporali.

Più in generale, tuttavia, gli interventi di maggiore necessità riguardano il completamento del caricamento dei testi già manualmente recuperati. Attualmente sono infatti presenti nel DAD solamente una parte, per quanto cospicua, di un solo canto (*Inf.* IX) dei tre preventivati. Ugualmente, si potrebbe arricchire ulteriormente anche la Sezione monografica. Inoltre, ad eccezione di poche schede, devono ancora essere inseriti i collegamenti ipertestuali e le note a piè di pagina interattive di cui si è opportunamente discusso nel corso di questo lavoro.

Si precisa infine che se queste e altre modifiche verranno applicate, sarà esclusivamente per favorire una migliore caratterizzazione della risorsa e sempre tenendo in considerazione le necessità degli utenti a cui essa si rivolge.

## **BIBLIOGRAFIA**

[1] Castellana, Riccardo. «Risorse digitali dantesche: testi, commenti, metrica, filologia». *Allegoria. Per uno studio materialistico della letteratura LXVIII* (2004): 96-124.

[2] Conway, Paul. «Digital Transformations and the Archival Nature of Surrogates». *Archival science XV* (2014): 51-69.

[3] Dollar, H. Charles. «La memoria elettronica e le ridefinizione della preservazione». *L'eclisse delle memorie*, a c. di Gregory, Tullio, e Morelli, Marcello. Bari, Laterza, 1994: 161-184.

[4] Irvine, Dean. «Editing Archives, Archiving Editions». *Journal of Canadian studies XL* (2006): 183-215.

[5] Maselli, Matteo. «Database Allegorico Dantesco». *Umanistica Digitale 13* (2022): 165-170.

[6] McCarty, Willard. *Humanities Computing*. Basingstoke-New York, Palgrave Macmillan, 2005.

[7] Price, Kenneth. «Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?». *Digital Humanities Quarterly III* (2009).

[8] Tomasi, Francesca. «Edizioni o archivi digitali? Knowledge site e apporti disciplinari». *Edizioni critiche digitali. Edizioni a confronto*, a c. di Bonsi, Claudia, e Italia, Paola. Roma, Sapienza Università Editrice, 2016.

# Tecnologie digitali per la realizzazione e adattamento di contenuti scolastici facilitati e semplificati

Maria Chiara Mattacchione<sup>1</sup>, Luca Del Greco<sup>2</sup>

<sup>1</sup> Liceo Scientifico 'P.Levi', Italia – m.mattacchione@outlook.it

<sup>2</sup> Saint Camillus International University, Italia – luca.delgreco@unicamillus.org

## ABSTRACT

L'articolo intende offrire una panoramica degli strumenti informatici e della loro applicazione nella realizzazione di materiale didattico semplificato e facilitato. Gli strumenti digitali costituiscono un grande supporto per la didattica, in particolare quando questa viene strutturata in ottica inclusiva. Grazie all'aumentata interattività, che permette allo studente di apprendere in un ambiente dinamico, sono numerose le potenzialità che le varie figure educative (docenti, tutor, adulti di riferimento) possono sfruttare padroneggiando le tecnologie. Le estese funzioni multimediali (come la possibilità di utilizzare grafiche, immagini, video o audio) permettono al docente di presentare le attività didattiche in una modalità nuova, facilitata rispetto alla didattica tradizionale. In questo articolo viene proposta una classificazione di tali tecnologie in base al loro utilizzo in ambito educativo. Le tre casistiche proposte sono corredate da altrettanti esempi applicativi, con l'obiettivo di fornire al docente spunti metodologici per pianificare una didattica inclusiva attraverso l'utilizzo delle tecnologie.

## PAROLE CHIAVE

Facilitazione; semplificazione; TIC; tecnologie; piattaforme digitali

## 1. INTRODUZIONE

Spesso, nella didattica ordinaria, in presenza di particolari difficoltà nell'apprendimento, il docente ha la necessità di avvalersi di materiale didattico adattato al fine di rendere l'esperienza scolastica partecipata e inclusiva.

Ci si riferisce alla facilitazione quando si compiono operazioni che mirano a rendere maggiormente accessibile un contenuto didattico, attraverso l'utilizzo di facilitatori come *flash-card*, giochi o prodotti multimediali, o attraverso interventi specifici come la riscrittura del testo. In questo caso si prevedono interventi, quali, ad esempio la modifica del layout, che comportano la scomposizione graduale delle difficoltà o una loro possibile attenuazione, senza però che le stesse siano eliminate in maniera consistente. Per semplificazione, invece, si intende l'operazione che porta alla riduzione consistente o addirittura all'eliminazione di parti significative del contenuto, intervenendo su elementi quali la sintassi o riducendo le richieste di un esercizio. [1] In tale contesto, gli strumenti digitali possono rappresentare un valido supporto per la realizzazione di una didattica adattativa. Le tecnologie permettono in molte situazioni di potenziare l'agire dell'insegnante: queste possono essere utilizzate sia come supporto per il docente nella realizzazione del materiale didattico facilitato e semplificato, sia come strumenti facilitanti (mappe, schemi e presentazioni) da somministrare all'alunno o da far realizzare direttamente all'alunno. Alcuni strumenti digitali, inoltre, possono costituire una valida piattaforma di interazione tra studenti e docenti attraverso quiz, giochi e condivisione di materiale scolastico.

## 2. CLASSIFICAZIONE

Nel presente lavoro si vuole proporre una classificazione delle casistiche relative ai diversi usi dello strumento tecnologico in ambito didattico, distinguendo tre macro-tipologie:

- **Caso a** Il docente si avvale degli strumenti digitali per realizzare contenuti da presentare all'allievo, prevedendo la realizzazione di materiale semplificato e facilitato (testi riscritti, con l'introduzione di un adeguato apparato iconografico) o di strumenti facilitanti (come ad esempio *flash-card*). In questo caso lo strumento digitale è funzionale al lavoro di predisposizione di materiale didattico adattato e pertanto non è direttamente proposto allo studente.
- **Caso b** L'alunno utilizza, in autonomia o supportato da una figura adulta, gli strumenti digitali. Tali strumenti possono essere utilizzati dallo studente per costruire supporti didattici, quali, per esempio, presentazioni Power Point attraverso le quali focalizzare meglio un argomento studiato in classe o supportare l'esposizione orale nel

momento della verifica. In questo caso lo strumento digitale è di ausilio allo studio individuale dell'alunno. La tecnologia è funzionale alla predisposizione del materiale ed è utilizzata direttamente dall'alunno.

· **Caso c** Il docente propone all'allievo attività su piattaforme digitali, quali, ad esempio, quiz per la verifica degli apprendimenti, piattaforme di condivisione dei contenuti e ulteriori stimoli interattivi. In questo caso non viene realizzato alcun prodotto, ma è lo stesso strumento digitale a fungere da ausilio allo studio individuale dell'alunno.

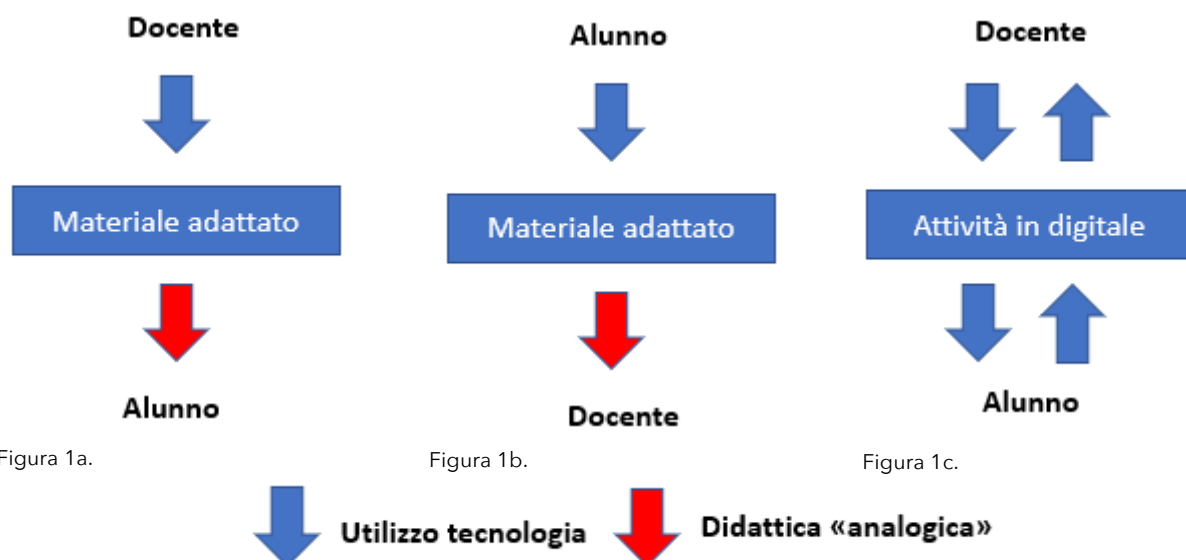


Figura 1. Schematizzazione delle modalità con cui il docente può agire sull'attività didattica attraverso l'uso della tecnologia. Figura 1a: il docente utilizza le tecnologie IT per realizzare del materiale da utilizzare in classe con lo studente. Figura 1b: lo studente utilizza le tecnologie IT per realizzare del materiale di supporto al proprio studio. Figura 1c: Il docente utilizza le tecnologie informatiche per predisporre e sottoporre allo studente un'attività in digitale.

Tale classificazione è di fondamentale importanza per l'attività del docente, che deve essere consapevole delle potenzialità e delle modalità con le quali utilizzare gli strumenti tecnologici. Di seguito sono illustrati alcuni esempi per ogni diversa casistica, nonché forniti ulteriori spunti di utilizzo relativi ad altri applicativi.

### 3. CASO A: UTILIZZO DI TECNOLOGIE DIGITALI DA PARTE DEL DOCENTE PER LA REALIZZAZIONE DI FACILITATORI E MATERIALE SEMPLIFICATO

Un primo utilizzo delle tecnologie digitali è rappresentato dall'applicazione delle stesse nella realizzazione di materiale didattico adattato. Per riscrivere un brano in forma facilitata o semplificata è possibile utilizzare numerosi software di elaborazione di testo e sfruttare le potenzialità della formattazione (intervenendo sulla dimensione di caratteri, sulla scelta di font adeguati e agendo su parametri quali l'interlinea). Il docente può rielaborare il testo in forma personalizzata e adattata e può inserire file multimediali. È possibile, inoltre, realizzare numerosi strumenti facilitati, utili per attuare una didattica adattativa nel contesto scolastico e a casa. A tale scopo hanno grande rilevanza le applicazioni di grafica web-based come Canva [2]. Con questo strumento è possibile, infatti, creare e personalizzare grafiche e infografiche di vario tipo in modo semplice e intuitivo, attraverso l'utilizzo di materiale preimpostato selezionabile dal catalogo di prodotti già fruibili dall'utente o alla possibilità di caricare e editare anche immagine private, inserendo intestazioni, sottotitoli o brani [3], come nell'esempio di Figura 2.



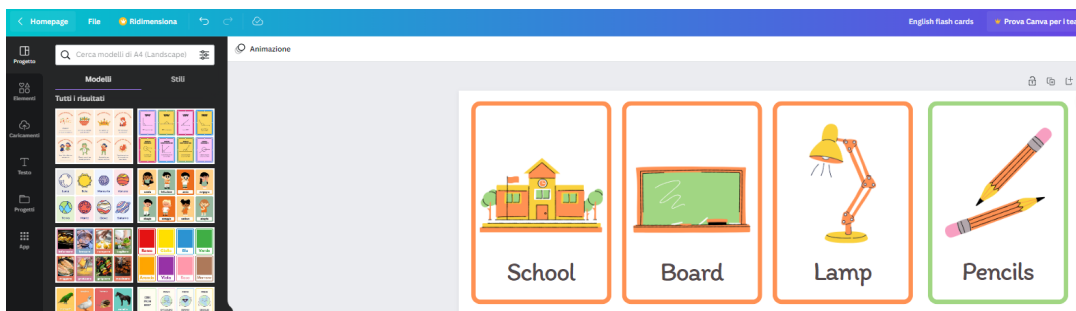


Figura 2. Utilizzo dell'applicativo Canva per la realizzazione di flash-card.

La Figura 3 mostra alcuni esempi di *flash-card* realizzate con Canva. Tali oggetti possono essere integrati nella didattica con finalità diverse: per aiutare lo studente nella costruzione di un glossario specifico, per la realizzazione di giochi di grammatica o per strutturare attività di *storytelling* con l'ausilio delle immagini. Nell'esempio specifico, sono mostrate *flash-card* funzionali all'apprendimento di alcuni termini in lingua inglese.

Il docente ha potuto realizzare le *flash-card* partendo da un formato già presente all'interno del catalogo di Canva. Tale prodotto è stato ulteriormente personalizzato attraverso l'introduzione di immagini afferenti al contesto scolastico (obiettivo di questa attività specifica è l'apprendimento di alcuni termini specifici relativi a questo ambito).

Ciascuna immagine è stata quindi corredata dal termine corrispondente in lingua inglese. Sono stati inoltre adottati ulteriori accorgimenti grafici (termini al singolare e termini al plurale sono contrassegnati con schede dal bordo di diverso colore).



Figura 3. English flash-card.

Le *flash-card* sono solo alcuni dei prodotti che possono essere realizzati con Canva o altri *tool* di grafica. Con lo stesso strumento possono essere realizzati anche giochi o schede. Tra i vari supporti facilitanti realizzabili attraverso gli applicativi digitali rientrano anche i *lapbook*, oggetti didattici, ibridi tra libro e quaderno, nei quali i contenuti (testi, immagini, grafici, linee del tempo, schemi, mappe concettuali, classificazioni e altro) sono presentati focalizzando l'attenzione sull'organizzazione concettuale e visiva delle informazioni [4]. Tali oggetti possono essere costruiti sia in forma cartacea -utilizzando *applicativi di grafica* per la progettazione- sia in forma digitale. A tal proposito è possibile utilizzare software come Adobe Indesign [5] per la realizzazione di pdf sfogliabili, pensati esattamente come veri e propri *lapbook* e utilizzabili dagli studenti in forma digitale. L'offerta di tale applicazione include anche la possibilità di inserire all'interno del libro espansioni multimediali come video o suoni.

#### 4. CASO B: UTILIZZO DI TECNOLOGIE DIGITALI DA PARTE DELLO STUDENTE PER LA PREDISPOSIZIONE DI MATERIALE DIGITALE O NON DIGITALE

La seconda casistica prevede che sia lo studente a realizzare del materiale attraverso l'impiego di tecnologie digitali. È bene evidenziare che tale attività può essere svolta autonomamente o con il supporto di un adulto di riferimento, da valutare sulla base delle competenze possedute dallo studente. Tra le attività afferenti a tale casistica sono sicuramente da menzionare gli strumenti di editing video, che possono consentire a studenti con difficoltà nella comunicazione di realizzare brevi clip che facilitino l'espressione anche attraverso l'aggiunta di parole, effetti ed elementi grafici. [6]. Costituiscono un ulteriore strumento di supporto gli applicativi per la realizzazione di presentazioni multimediali, come Power Point [7]. Con tali software lo studente può costruire in autonomia elaborati, digitali o stampabili, da utilizzare per lo studio, il ripasso

o l'esposizione orale. Molti strumenti di facilitazione costituiscono un valido supporto per lo studente nelle attività di elaborazione di un discorso o di comprensione e interpretazione di un racconto. La narrazione, infatti, è un'attività di grande importanza per il percorso di apprendimento: questa consente di assimilare scenari e modelli di vita e di trasferirli sul piano immaginativo e fantastico; la narrazione può inoltre essere utile per sviluppare attenzione, autocontrollo, immaginazione e per avviare altre attività didattiche. [8]. L'attrazione che gli studenti hanno per la tecnologia -sia per il carattere di novità sia per l'affinità a strumenti da loro utilizzati quotidianamente- può essere sfruttata per stimolare la creatività dell'allievo nell'ambito della narrazione. Un genere che assume grande interesse ed efficacia in ambito didattico è il fumetto. Proporre un'attività di lettura di un fumetto può essere molto utile, allo stesso modo può esserlo la proposta di un'attività di produzione [9]. Un software online e gratuito utilizzabile per tale attività è Storyboard That [10]: un portale all'interno del quale è possibile creare facilmente le proprie vignette scegliendo tra sfondi relativi a vari contesti. Una volta inserito lo sfondo è possibile aggiungere anche personaggi, oggetti e testi. [11]. Nell'esempio riportato in Figura 4, lo strumento Storyboard That viene proposto allo studente per la realizzazione di un riassunto, sotto forma di fumetto, inerente a un episodio epico affrontato in classe. Lo studente ha a disposizione diverse istantanee nelle quali collocare le scene del racconto (Fig. 4a); può inoltre inserire le diverse ambientazioni e definire i personaggi, personalizzando abiti e caratteristiche fisiche, intervenendo sulle pose e modificando le espressioni del volto in funzione dello stato d'animo (Fig. 4b). Possono poi essere inseriti i fumetti, all'interno dei quali collocare il testo (Fig. 4c).

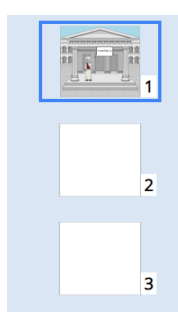


Figura 4a.

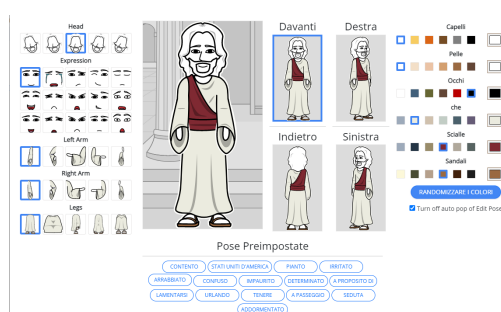


Figura 4b.

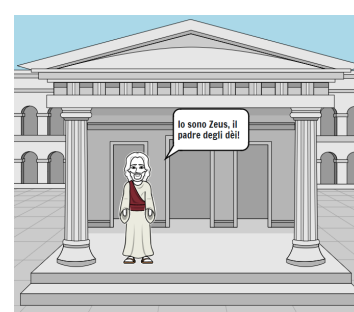


Figura 4c.

Figura 4. Realizzazione di un fumetto con Storyboard That.

## 5. CASO C: SOMMINISTRAZIONE DA PARTE DEL DOCENTE DI UN'ATTIVITA' FACILITATA O SEMPLIFICATA IN UNA PIATTAFORMA DIGITALE

L'emergenza da Covid-19 e le conseguenti misure di contenimento della pandemia hanno comportato modifiche significative anche in relazione alle modalità di erogazione e svolgimento della didattica. Nel periodo pandemico, a causa della necessità di organizzare lezioni in didattica a distanza (DaD) e didattica digitale integrata (DDI), docenti e alunni si sono avvalsi dell'uso di numerose piattaforme per l'organizzazione, lo svolgimento e la condivisione delle attività didattiche. Tra le piattaforme più diffuse si annoverano: Google Classroom [12], diffusasi rapidamente grazie all'importante integrazione con le diverse app di Google all'interno dello spazio di lavoro di Google Workspace for Education [13], e Microsoft Teams [14], accessibile attraverso il servizio web-based fornito da Microsoft, tramite la piattaforma online Office 365, specificamente pensata per il mondo della scuola, in cui sono messe a disposizione anche le ulteriori risorse della suite Office. A margine si evidenzia anche la presenza della piattaforma Schoolwork [15] pensata da Apple per aiutare il docente a condividere file con gli studenti, monitorare i progressi, dare riscontri e proporre attività da salvare nell'archivio virtuale. [16]. Tali piattaforme sono entrate profondamente nella quotidianità non solo della scuola, ma di numerosi ambiti lavorativi. L'acquisizione di una maggiore consapevolezza sul loro utilizzo sta senza dubbio aiutando la comunità educante nell'indagine di ulteriori, nuove applicazioni anche in ottica inclusiva. A tal proposito si può ipotizzare l'utilizzo di tali software anche in contesti post-pandemici, dove la didattica è organizzata di norma in presenza, ma in cui soluzioni integrate possono andare a colmare specifici bisogni. A titolo di esempio si cita la possibilità, per studenti che vivono in posti lontani o sono impossibilitati a incontrarsi, di organizzare sessioni di studio o lavoro di gruppo in orario extrascolastico, aumentando sia l'interazione all'interno del gruppo dei pari che l'efficacia dello studio cooperativo.

In aggiunta alle piattaforme già annoverate possiamo annoverare numerosi esempi di applicativi specifici, generalmente disponibili online, che consentono al docente di impostare un'attività didattica fruibile direttamente in forma digitale, in chiave semplificata o facilitata. In Figura 5 è riportato un esempio. Attraverso la piattaforma Kahoot [17], il docente può pianificare verifiche, giochi ed esercitazioni in forma digitale. Nell'esempio specifico è riportata la realizzazione di un quiz

di Storia dell'arte. In questo caso gli elementi facilitati che la piattaforma consente di inserire rispetto a un'attività didattica di tipo tradizionale sono la possibilità di caricare supporti multimediali (implementando, per esempio, l'apparato iconografico, molto spesso non presente nel materiale didattico cartaceo o a volte – se stampato - privo di colori o di una risoluzione adeguata). Ulteriori peculiarità di questo applicativo sono l'elevata interattività e la presenza dell'elemento ludico: tale applicazione, infatti, può essere utilizzata dal proprio device (smartphone o tablet) e permette di giocare in squadra, vedere statistiche e classifiche fra compagni in tempo reale) .

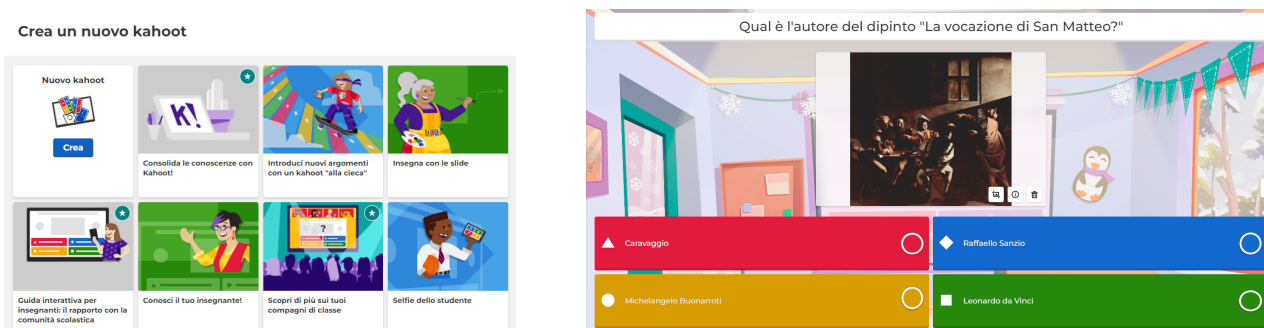


Figura 5. Utilizzo di Kahoot per la realizzazione di un quiz con la presenza di elementi multimediali.

## 6. CONCLUSIONI

Con il presente contributo si è voluto proporre una tassonomia in grado di classificare i diversi utilizzi delle tecnologie in ambito didattico, distinguendo tre categorie specifiche, corredate da altrettanti esempi esplicativi. I contenuti mostrati non sono comprensivi dell'intero panorama disponibile, ma consentono di evidenziare le differenze tra le categorie proposte. Aver definito una classificazione in tal senso può essere d'aiuto per i docenti, al fine di renderli maggiormente consapevoli circa la possibilità di applicazione dei vari strumenti all'interno del contesto scolastico, sia per gli sviluppatori che possono in tal modo orientare lo sviluppo dei software anche in funzione delle esigenze della comunità scolastica.

## BIBLIOGRAFIA

- [1] Cottini, Lucio, e Mattacchione Maria Chiara. Semplificare e facilitare i contenuti didattici, gli esercizi e le verifiche: come e cosa fare. Giunti EDU, 2022
- [2] Canva, [www.canva.com](http://www.canva.com) (consultato online il 15/02/2023)
- [3] Ferrarese, Gianni. Infografiche con Canva. 101 idee per una didattica digitale integrata. Erickson, 2021
- [4] Frigo, Maria. Sesamo didattica interculturale, Giunti Scuola, 2017
- [5] Adobe Indesign, [https://it.wikipedia.org/wiki/Adobe\\_InDesign](https://it.wikipedia.org/wiki/Adobe_InDesign) (consultato online il 15/02/2023)
- [6] Minervini Magda, Percorsi di inclusione nella didattica digitale. 101 idee per una didattica digitale integrata. Erickson, 2021
- [7] Power Point, <https://www.microsoft.com/it-it/microsoft-365/powerpoint> (consultato online il 15/02/2023)
- [8] Calvani Antonio, Marzano Antonio Marzano e Morganti Annalisa. La didattica in classe. Casi, problemi e soluzioni. Carocci editore, 2021
- [9] Bianchi, Ferruccio e Farello Patrizia. Lavorare sul fumetto. Unità didattiche e schede operative. Erickson,2000
- [10] Storyboard that, [www.storyboardthat.com](http://www.storyboardthat.com) (consultato online il 15/02/2023)
- [11] Tonioli, Davide. Fumetti digitali con Storyboard that. 101 idee per una didattica digitale integrata. Erickson, 2021
- [12] Classroom, <https://classroom.google.com/> (consultato online il 15/02/2023)
- [13] Chiara, Spalatro. Assegnare e correggere compiti con Google Classroom. 101 idee per una didattica digitale integrata. Erickson, 2021
- [14] Microsoft Teams, <https://www.microsoft.com/it-it/microsoft-teams/group-chat-software> (consultato online il 15/02/2023)
- [15] Apple Schoolwork, <https://apps.apple.com/it/app/schoolwork/> (consultato online il 15/02/2023)
- [16] Parise Samantha. Schoolwork, la piattaforma Apple. 101 idee per una didattica digitale integrata. Erickson, 2021
- [17] Kahoot <https://kahoot.it/> (consultato online il 15/02/2023)

## APPENDICE

### Credits immagini

Figura 1. © Mattacchione Maria Chiara e Del Greco Luca

Figure 2-3. © Mattacchione Maria Chiara tramite Canva.com

Figura 4. © Mattacchione Maria Chiara tramite Storyboard That

Figura 5. © Mattacchione Maria Chiara tramite Kahoot

# **Proteus: marcatura di testi per una lettura antropologica del mito greco e romano**

Gloria Mugelli<sup>1</sup>, Cristiana Franco<sup>2</sup>

<sup>1</sup> Ricercatrice indipendente, Italia – gloria.mugelli@gmail.com

<sup>2</sup> Università per Stranieri di Siena, Italia – franco@unistrasi.it

## **ABSTRACT**

Il contributo illustra le varie fasi di costruzione di una piattaforma web in cui si presenta una lettura antropologica di alcuni episodi del mito greco e romano. I testi dei racconti mitici, tratti da saggi scritti da esperti di antropologia del mondo antico, sono stati marcati con i nomi di personaggi e luoghi, e con un'annotazione tematica. È in fase di costruzione un'ontologia, per la strutturazione e l'interrogazione dei dati annotati.

## **PAROLE CHIAVE**

Mitologia; annotazione tematica stand-off; ontologie; antropologia del mondo antico

## **1. INTRODUZIONE**

La piattaforma web *Proteus* propone una lettura antropologica di alcuni episodi del mito greco e romano. Il progetto nasce da un lavoro collettivo sul mito svolto da antropologi del mondo greco e romano<sup>1</sup>, che ha come esito il volume *Il sapere mitico* [1], che riflette su grandi categorie antropologiche a partire dai racconti mitici, intesi non soltanto nel loro valore narrativo, ma anche come strumenti di trasmissione di valori, codici di comportamento, forme di pensiero. Il volume è diretto sia agli studiosi del mondo antico, interessati ad una prospettiva antropologica sul mito, ma è anche rivolto ad un pubblico non esclusivamente specialistico, ed è pensato come strumento di conoscenza, oltre che di approfondimento, del mito e dell'antropologia del mondo antico.

Il progetto *Proteus* nasce dall'esigenza di rendere accessibili sul web parte dei contenuti prodotti dal lavoro svolto per *Il sapere mitico*. È infatti a partire dalle fasi finali di redazione del volume edito per Einaudi che si avverte l'esigenza di non mettere da parte l'enorme mole dei materiali preparatori redatti dagli autori. *Proteus* si propone, quindi, non tanto come un progetto parallelo e indipendente dal lavoro editoriale, realizzato da zero per una pubblicazione sul web, ma come un prodotto generato dall'officina de *Il sapere mitico*, attraverso una ristrutturazione dei testi e dei materiali che permetta agli utenti finali di avere un'esperienza diversa di approccio ai testi e ai temi proposti dal gruppo di ricerca, complementare a quella della lettura del volume.

L'obiettivo del progetto è infatti quello di dare continuità alla riflessione portata avanti nel volume, mettendo a frutto il metodo dell'antropologia del mondo antico per offrire all'utente della piattaforma web un'esperienza personalizzabile di approccio al mito.

Il processo messo in atto per la realizzazione della piattaforma, illustrato in questo contributo, mostra come, grazie alla mediazione delle *digital humanities* e grazie all'uso di accorgimenti poco dispendiosi in termini di tempo per gli umanisti, sia possibile estrarre da testi di ricerca e divulgazione una significativa quantità di dati, che permettono di ristrutturare le informazioni e interrogarle in modi nuovi e originali.

## **2. L'OFFICINA DI SAPERE MITICO E I MATERIALI**

La sfida del progetto *Proteus* è quella di creare un prodotto digitale a partire dai risultati di un lavoro di ricerca svolto da esperti di antropologia del mondo greco e romano secondo metodi del tutto tradizionali.

Il lavoro per la costruzione della piattaforma si è sviluppato a partire da un testo molto strutturato: si sono infatti recuperate le versioni preparatorie integrali dei saggi redatti dai vari autori. I testi seguono, in linea generale, la scansione proposta nell'indice del volume: le diverse sezioni trattano gli argomenti cardine per una lettura antropologica del mito antico (l'essere umano, l'individuo e la collettività, il genere, gli animali, il divino, la parentela, l'economia, la salute e la malattia, lo spazio e il tempo, la comunicazione), e ogni sezione è divisa in due parti, che affrontano il tema nel mito greco e in quello romano.

---

<sup>1</sup> Il lavoro su *Il sapere mitico* e quello per il progetto *Proteus* sono stati finanziati da due progetti PRIN del MIUR.

Ogni saggio affronta sotto diversi aspetti uno dei temi portanti per la civiltà greca e romana, adottando un'ottica antropologica, e riflette su come dai racconti mitici si possono ricostruire le strutture fondamentali del pensiero antico. Per supportare l'argomentazione vengono citati diversi racconti mitici, i più importanti dei quali sono riassunti dagli autori, in modo che sia possibile seguire, oltre alla trattazione antropologica, anche la narrativa del mito che si sta analizzando. Si è scelto di lavorare su questi racconti sintetici dei miti, per due ragioni: se da una parte i riassunti del mito permettono di seguire l'ossatura del ragionamento degli autori, dall'altra questi funzionano come unità che possono essere messe in relazione l'una con l'altra, al di là del filo conduttore della trattazione del volume, per creare nuove connessioni e nuove ipotesi di lettura e di interpretazione.

Sfruttando l'impaginazione del testo di partenza, che presentava i racconti mitici con un rientro rispetto al saggio in cui erano inseriti, il testo dei racconti è stato estratto e codificato in XML-TEI. A tutti i racconti sono stati assegnati due identificatori: il primo, numerico, che li riconducesse alla loro posizione nel testo di partenza (sezione e capitolo), e un'etichetta che facesse riferimento all'episodio mitico riassunto nel testo.

### 3. ENTITÀ NOMINATE

Una volta isolati i materiali su cui lavorare per la costruzione di *Proteus*, si è lavorato per estrarre dal testo la maggiore quantità possibile di informazioni. Trattandosi di episodi del mito, uno dei dati senza dubbio più rilevanti è quello dei nomi dei personaggi in gioco nel racconto. Come prima cosa, sono state estratte le entità nominate (nomi di persona e luoghi) utilizzando il sistema messo a disposizione dalla piattaforma T2K<sup>2</sup>, sviluppata dall'ILC-CNR di Pisa [2].

Una volta estratti nomi di luoghi e personaggi, questi sono stati marcati nel testo XML di partenza con i due tag <name> e <placename>. Trattandosi di una quantità di dati limitata, è stata svolta a mano una disambiguazione dei casi dubbi come, per esempio, i casi in cui il termine "Ade" si riferiva al dio degli Inferi, e quelli in cui il termine si riferiva agli Inferi come luogo. Lavorando sulla lista delle entità nominate estratte dal testo attraverso OpenRefine [3], si è poi associato ad ogni entità prima l'equivalente latino e successivamente l'ID univoco del personaggio mitico presente su Wikidata. Con l'attribuzione dell'ID wikidata, si sono potuti risolvere anche i casi di ambiguità in cui un personaggio era citato con nomi diversi, in particolare nei casi di personaggi romani citati con uno, con due o con tutti e tre i loro nomi.

```

1 <div3 n="176" xml:id="nascita_minotauro">
2 <head>Nascita del minotauro</head>
3 Mentre <name>Minosse</name>, aspirante al trono di <placeName>Creta</placeName>, offriva un sacrificio a
4 <name>Poseidone</name>, per dimostrare ai propri concittadini il proprio legame speciale con il dio gli chiese di
5 far emergere un toro dal mare e promise che glielo avrebbe sacrificato. <name>Poseidone</name> fece allora
6 apparire per lui un toro bellissimo, ma <name>Minosse</name> mandò il toro tra le sue mandrie e ne sacrificò un
7 altro. Riuscì così a diventare re di <placeName>Creta</placeName> e fu il primo ad avere il dominio sul mare (la
8 "talassocrazia"), ma il dio, adirato con lui, rese furioso l'animale e fece in modo che <name>Pasifae</name> se
9 ne invaghisse. Innamorata del toro, la regina si fa aiutare dall'ingegnoso <name>Dedalo</name>, fuggito da
10 <placeName>Atene</placeName> a causa di un omicidio.
11 <name>Dedalo</name> costruì una vacca di legno e la pose su delle ruote, dentro la fece cava, le cucl addosso la
12 pelle di una vacca che aveva scuoiato, la collocò nel prato dove il toro era solito pascolare e vi fece salire
13 <name>Pasifae</name>. Sopraggiunse il toro che si unì a lei come se fosse realmente una vacca. Dopo questa
14 unione, la regina generò <name>Asterio</name>, detto <name>Minotauro</name>, che aveva la testa di toro e il
15 corpo di uomo. In seguito a certi oracoli, <name>Minosse</name> lo rinchiuso, ben custodito, nel celebre
16 labirinto, un edificio costruito da <name>Dedalo</name> che, con i suoi tortuosi corridoi, impediva di trovare
17 l'uscita (<bibl>Apollodoro, Bibl. 3, 1, 3-4</bibl>).
18 </div3>

```

Figura 1. Esempio di testo di un racconto mitico.

### 4. ANNOTAZIONE TEMATICA STAND-OFF

Le informazioni più significative per quanto riguarda i temi antropologici presenti nel racconto mitico non si trovano tanto nel testo dei riassunti, quanto nel contesto del saggio in cui questi sono citati. Per questo è stata svolta un'annotazione tematica stand-off, in cui si fanno corrispondere all'ID del racconto mitico una serie di parole chiave che marcano fenomeni interessanti dal punto di vista antropologico, oltre a elementi particolari che emergono nel mito.

Il tagset utilizzato per l'annotazione dei testi è stato costruito dagli autori dei vari contributi, che hanno identificato alcuni dei temi chiave fondamentali per la lettura dei loro testi, confrontandosi poi con gli altri autori in modo da creare un set unico, strutturato per macrocategorie e singoli tag.

L'annotazione è la parte del lavoro in cui si dà quindi spazio al lavoro interpretativo svolto dagli autori sui diversi miti, e in cui si riflette il taglio antropologico dato alla lettura dei racconti.

176	nascita_minotauro	mostro; minotauro; nascita; toro; sacrificio; animali_prodigiosi; castigo_divino; artigianato; simulacro; oracolo; labirinto
-----	-------------------	--

Figura 2. Esempio di annotazione di un episodio mitico.

L'annotazione svolta sul racconto della nascita del minotauro è un esempio di come questo mito, inserito nel capitolo sui mostri e gli ibridi nella mitologia greca, si possa connettere a temi fondamentali trattati nei capitoli sulla ritualità (il sacrificio), sugli animali e il loro rapporto con gli umani, sul rapporto tra uomini e dèi, sulle arti e l'artigianato.

Allo stato attuale i testi, corredati di marcatura delle entità nominate e annotazione tematica, andranno a costruire un primo prototipo della piattaforma *Proteus*, atteso nel corso del 2023. Nel prototipo, sarà possibile scorrere i testi in maniera consequenziale, seguendo la struttura del volume *Sapere mitico*; i testi saranno accompagnati, oltre che dall'annotazione tematica, anche da un breve commento antropologico redatto dagli autori.

Il sito potrà essere consultato anche sfruttando la marcatura delle entità nominate, leggendo tutti i racconti in cui è citato un determinato personaggio, oppure seguendo l'annotazione tematica per connettere tutti i racconti legati ad un particolare tema antropologico.

## 5. COSTRUZIONE DELL'ONTOLOGIA

Si sta attualmente lavorando per potenziare la possibilità di ricerca dei diversi tipi di dati estratti o annotati sui testi, integrando i dati presenti e ricercabili su Wikidata con l'annotazione tematica. Seguendo il modello di altri progetti ([4];[5]), si sta realizzando un'ontologia dei tag utilizzati per l'annotazione.

L'ontologia adotta come *upper level* DOLCE Lite (<http://www.loa.istc.cnr.it/dolce/overview.html>) [6], e segue quindi l'approccio adottato per il progetto *Memorata Poetis* [4], dal quale è stato ripreso interamente anche il *mid level*, che permette di agganciare DOLCE all'ontologia di dominio. L'ontologia di *Proteus* condivide quindi moltissime classi con il progetto *Memorata Poetis*; la vicinanza delle due ontologie permetterà, una volta implementato un sistema di ricerca che permetta di interrogare il database *Proteus*, di osservare anche come i temi osservati nello studio antropologico dei miti siano affrontati dai testi poetici annotati per il progetto *Memorata Poetis*.

Per i nomi dei personaggi, i luoghi, le categorie meno specifiche del dominio, l'ontologia rimanda all'ID Wikidata. Inoltre, seguendo lo stesso approccio, sia i nomi di personaggi e luoghi, sia le entità corrispondenti ai singoli eventi mitici, potranno essere messi in relazione anche con gli URI del progetto MANTO di Mythlab (<https://www.manto-myth.org/manto>) [7], per il quale è stato creato un enorme dataset mitologico, basato su fonti letterarie.

L'integrazione degli identificatori di MANTO e Wikidata ha lo scopo di mantenere separati, nell'ontologia creata per il progetto *Proteus*, due aspetti che inevitabilmente si fondono quando si ha a che fare con una raccolta di miti: *Proteus* non è infatti un catalogo o un'enciclopedia del mito, ma si vuole proporre come strumento ermeneutico. Per questo motivo, le risorse esterne verranno sfruttate per non tentare inutilmente di attribuire alla piattaforma *Proteus* una completezza enciclopedica che i testi schedati non hanno.

La principale funzione dell'ontologia *Proteus* sarà infatti quella di connettere gli eventi dei singoli racconti mitici tramite diverse proprietà alle categorie chiamate in causa nell'annotazione tematica, che riflette il lavoro ermeneutico svolto dagli antropologi del mondo antico.

Queste proprietà dell'ontologia saranno quelle più rilevanti dal punto di vista dell'interesse specifico del lavoro condotto su *Il sapere mitico*, perché permetteranno di connettere un evento o un personaggio a tutte le categorie rilevanti per lo studio del mito e dell'antropologia del mondo antico.

Per esempio, il minotauro di Creta ha una sua entry Wikidata (Q129866), come istanza della classe degli ibridi uomo-animale, in cui compaiono moltissimi dettagli, dalle parentele all'aspetto fisico del mostro. Per il testo del racconto, sarà quindi sufficiente specificare che questo ha come personaggio il minotauro. Tramite una proprietà dell'ontologia, sarà poi specificato che il racconto chiama in causa diverse categorie significative dal punto di vista antropologico, emerse nell'annotazione tematica.

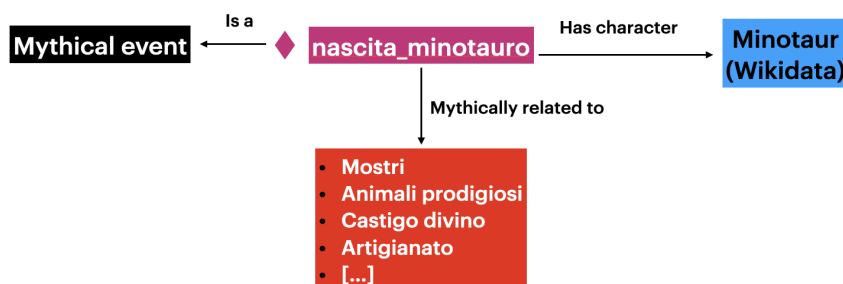


Figura 3. Esempio di proprietà impiegate nell'ontologia.



## 6. CONCLUSIONI

Una volta ultimata l'ontologia, verrà costruito un sistema di interrogazione che, attraverso la costruzione di *queries* SPARQL, permetterà di interrogare sia l'ontologia di dominio costruita per il progetto, sia i dati presenti su altre risorse. Questo sistema di interrogazione avanzata andrà a integrare un sistema di ricerca semplice, basato sulla co-occorrenza di nomi e parole chiave nell'annotazione.

Integrata con i due sistemi di ricerca, la piattaforma Proteus offrirà uno strumento di lettura a più livelli di alcuni episodi del mito greco e romano: da una parte, sarà possibile una lettura consequenziale dei racconti mitici, brevemente commentati, che potrà dare una conoscenza generale del tema e una chiave d'accesso per una lettura antropologica del mito antico, andando a costruire una sorta di strumento di integrativo al saggio *Il sapere mitico*, di cui i racconti costituiscono l'ossatura.

Dall'altra parte, la piattaforma offrirà a chi lo desidera la possibilità di interrogare i materiali elaborati dagli autori del volume, andando così a costruire percorsi nuovi all'interno delle diverse sezioni del libro e delle diverse tematiche affrontate, in modo da mettere in relazione diversi tipi di informazione e creare nuovi spunti per la lettura e l'analisi dei racconti mitici annotati presenti sulla piattaforma Proteus.

In conclusione, se dal punto di vista della produzione dei dati questo progetto non costituisce un'innovazione, ma anzi si inserisce su un terreno già battuto per esempio dal progetto MANTO, dal punto di vista della costruzione del sistema e del flusso di lavoro il progetto Proteus può costituire un esempio positivo di come, con alcuni accorgimenti e con il supporto di esperti di Digital Humanities, un lavoro svolto secondo metodi tradizionali da specialisti di dominio possa essere rimodellato per la costruzione di una piattaforma digitale che renda fruibile il sapere prodotto con diversi metodi, per diversi tipi di pubblico.

## BIBLIOGRAFIA

- [1] Bettini, Maurizio, et al. a c. di. *Il sapere mitico. Un'antropologia del mondo antico*. Einaudi, 2021.
- [2] Dell'Orletta, Felice, Giulia Venturi, Andrea Cimino, e Simonetta Montemagni. «T2K2: A System for Automatically Extracting and Organizing Knowledge from Texts». In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, 2062–70. Reykjavik, Iceland, 2014.
- [3] Hooland, Seth van, Ruben Verborgh, e Max De Wilde. «Cleaning Data with OpenRefine». *Programming Historian*, 5 agosto 2013. <https://programminghistorian.org/en/lessons/cleaning-data-with-openrefine>.
- [4] Kahn, Fahad, Silvia Arrigoni, Federico Boschetti, e Francesca Frontini. «Restructuring a Taxonomy of Literary Themes and Motifs for More Efficient Querying». *Estudos Literários Digitais* 2 4, fasc. 2 (2016): 11–27.
- [5] Mugelli, Gloria, Andrea Bellandi, Federico Boschetti, e Anas Fahad Khan. «Designing an Ontology for the Study of Ritual in Ancient Greek Tragedy». In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*. Montpellier: Association for Computational Linguistics, 2017. <http://aclweb.org/anthology/W17-7011>
- [6] Gangemi, Aldo et al., «Sweetening Ontologies with DOLCE», in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, a c. di Asunción Gómez-Pérez e V. Richard Benjamins, *Lecture Notes in Computer Science* (Berlin, Heidelberg: Springer, 2002), 166–81, [https://doi.org/10.1007/3-540-45810-7\\_18](https://doi.org/10.1007/3-540-45810-7_18)
- [7] «A Dataset of Mythical People with Stable URIs», MYTHLAB, consultato 5 maggio 2023 <https://www.manto-myth.org/blog/a-dataset-of-mythical-names-with-stable-uris>[https://doi.org/10.1007/3-540-45810-7\\_18](https://doi.org/10.1007/3-540-45810-7_18)
- [8] Mugelli, Gloria, Federico Boschetti, Riccardo Del Gratta, Angelo Mario Del Grosso, Fahad Kahn, e Andrea Taddei. «A user-centred design to annotate ritual facts in ancient greek tragedies». *BICS* 59, fasc. 2 (2016): 103–20.

# Seeing the Unseen: an approach for visualising situated statements in the Cultural Heritage domain

Carlo Teo Pedretti<sup>1</sup>, Andrea Schimmenti<sup>2</sup>, Francesca Tomasi<sup>3</sup>, Fabio Vitali<sup>4</sup>

<sup>1</sup> University of Bologna, Italy – carlo.teo.pedretti2@unibo.it

<sup>2</sup> University of Bologna, Italy – andrea.schimmenti2@unibo.it

<sup>3</sup> University of Bologna, Italy – francesca.tomasi@unibo.it

<sup>4</sup> University of Bologna, Italy – fabio.vitali@unibo.it

## ABSTRACT

Data visualisation tools, including semantic technologies, help users to explore novel connections in a dataset and generate new knowledge. In recent years, semantic models have been the subject of new formalisations capable of effectively conveying complex qualitative information, including meta-information (provenance, reliability) and contexts, particularly relevant in the cultural heritage (CH) domain where historical data are often ambiguous, contested, or missing. Nevertheless, there is still a lack of adequate insights into the practices and epistemological models in visualising uncertain data for the humanities. The present paper, therefore, aims to: (1) identify some of the main strategies for representing and visualizing qualitative information in the digital humanities domain; and (2) suggest the use of higraphs as a technique for visualizing complex statements in the context of semantic applications.

## KEYWORDS

Data visualization; higraph; visual epistemology; uncertainty; UX

## 1. INTRODUCTION

The joint effort undertaken by numerous cultural institutions in the Libraries, Archives and Museums (LAM) circuit to digitize and make their corpora available in digital format has resulted in an unprecedented availability of data from cultural heritage. In parallel, the adoption of semantic web principles (such as Linked Open Data) has enabled the semantic enrichment of heterogeneous resources, ensuring their interoperability with a view to knowledge transmission and querying through inferential processes [12].

Thus arises the need to rethink presentational level and user experience (UX) of semantic portals in cultural contexts. Specifically, by UX, we mean a multidimensional concept that extends and incorporates the dimensions of usability and accessibility, as well as all those pragmatic and hedonic factors that contribute to end-user satisfaction before, during, and after interaction with a system [6]. A crucial point in the field of UX research is related to the dimension of user interfaces, as they are constitutive of the presentational dimension through which most user interactions with collection data occur. Well-established paradigms in user interface (UI) design for digital cultural collections include the use of generous interfaces, which allow switching between synoptic and idiographic views, while most common approaches to information visualisation involve the use of multidimensional representations of data (quantitative, ordinal, categorical, or textual) where graphs, maps, and timelines represent some of the most commonly used abstract models [14, 15].

However, in the humanities, where historical data are often ambiguous or biased and hermeneutic practice is a constitutive element of epistemic processes, uncertainty<sup>1</sup> plays a crucial role in the domain; therefore, it becomes a common practice to include both preferred and disputed (or deprecated) assertions [1].

Conversely, data visualisation always involves a process of more or less explicit abstraction and parameterisation of uncertain phenomena, resulting in their removal, adjustment or marginalisation. Thus, one can see how the choice of a visualisation paradigm directly impacts the end user's ability to discern complex information correctly. In addition, when dealing with contested or ambiguous data, the omission of uncertainty at the visualisation stage due to inadequate interfaces or biased representations, is likely to foster erroneous or uninformed interpretations of the available data [11]. This is all the more problematic considering that most visualization strategies used in the CH landscape are incorporated from disciplines whose epistemological premises differ radically from those typical of the humanities.

---

<sup>1</sup> We define by the term *ontological uncertainty* a structural property of an entity. This phenomenon is independent of the degree of knowledge we may have about that given entity. In contrast, we define by the term *epistemic uncertainty* a phenomenon which affects the interpretive and cognitive process of a subject around a particular event, and thus primarily concerns the perceiver [9]. In the present work we refer primarily to the second meaning of uncertainty.

Such practices fall within what [4] defines under the category of *representational paradigm*. This paradigm, as we have seen, results in the concealment of the interpretive dimension that qualitative data require, in favor of simplified views that are often unable to restore the ambiguity and opacity typical of the objects they represent [4, 15]. In parallel, these paradigms surreptitiously re-propose a realist model of knowledge where, ultimately, the results of the visualisation appear ontologically adequate to the very phenomena they denote.

On the contrary, by introducing uncertainty and contextuality into the data modelling process, it becomes evident that the data displayed are not natural representations of facts but remnants of a reductionist process. Emphasizing that data are always interpreted, situated and constructed from a perspective, the *non-representational paradigm* proposed by [3] employs the term *capta* for defining the output of this data remediation.

The purpose of this research is thus oriented by the need to explore emerging approaches in the field of visualisation of uncertain semantic data: on the one hand, to reaffirm the inherently critical dimension of digital humanities and its methodologies; on the other hand, to investigate and propose models of visual epistemology appropriate to the interpretive practices in use in historical disciplines.

## 2. MAPPING SITUATED AND SITUATING STATEMENTS

In this regard, in addition to traditional solutions offered by semantic technologies for modelling complex information (e.g., context, provenance, etc.), such as reification (describing a statement with another statement), N-ary relations (statements about the relationship about two entities), and Named Graphs, fruitful proposals have experimented in the field of *digital hermeneutics* [2] through the creation of formal models and ontologies such as RDF-star<sup>2</sup>, HiCO<sup>3</sup> and uncertainty-supporting extensions for CIDOC-CRM<sup>4</sup> [13].

Nonetheless, complex statements concerning the interpretation of a phenomenon often coexist with other conflicting statements in an evolving critical space. While time may let one particular statement emerge among others, it is important for a scholar to take into account (or visualise) all the previous (deprecated) statements to be able to produce new interpretations adequately. A statement is always a statement *about* something. This aboutness or *situatedness* can be articulated in terms of a hierarchical structure where a *situated statement* (e.g., a statement about something) is referred to and enclosed by a *situating statement* (a meta-statement about the previous statement), defining a *situation* (e.g., a complex statement). Accordingly, it is possible to define three levels concerning the architectonic of complex statements:

- *data level*, composed of atomic facts or phenomena.
- *inference level* (or *situated level*), created when relating elements are taken from the data level.
- *meta-inference level* (or *situating level*), constructed when producing a statement about a given inference.

Consequently, a visualisation strategy capable of adequately representing complex statements must be able to address these three levels through an appropriate mapping between the logic structure and.

## 3. VISUALIZING COMPLEX STATEMENTS USING HIGRAPH

In the context of Linked Data (LD) collections, data visualisation is considered a crucial epistemic aid [8] that can foster the discovery of novel relationships between data and the constitution of new knowledge. Some of the most common RDF data visualisation metaphors use graphs or tables. Most online tools support graph views, while most knowledge bases use tabular views [10]. The graph usually provides a synoptic view of the collection and is an appropriate tool for discerning clusters of related data.

Nevertheless, mature tools capable of visualising these formal models are not numerous. Specifically, GraphDB<sup>5</sup> is a largely employed graph database with RDF and SPARQL support that supports reification, N-ary relations, Named Graphs, and RDF-star, as well as also providing a graph visualization tool.

For the purpose of explanatory clarity, we present an example containing some assertions around the creator (*wdt:P170*) of the painting *Vue de Marseille: Joute et fête sur l'eau* (*wd:Q3563742*), taken from Wikidata<sup>6</sup>. Wikidata represents uncertainty by using the property “stated in” (P248). This property describes the process of giving credit for the statement

<sup>2</sup> [https://w3c.github.io/rdf-star/cg-spec/editors\\_draft.html](https://w3c.github.io/rdf-star/cg-spec/editors_draft.html).

<sup>3</sup> <https://marilenadaquino.github.io/hico/>.

<sup>4</sup> <https://www.cidoc-crm.org/Version/version-6.2>.

<sup>5</sup> <https://graphdb.ontotext.com/>.

<sup>6</sup> [https://www.wikidata.org/wiki/Wikidata:Main\\_Page](https://www.wikidata.org/wiki/Wikidata:Main_Page).

to something or someone. Using the property `wikibase:Rank` in case of multiple statements, it specifies the best one based on bibliography. ,

The first assertion or *situating statement* is produced by (`pr:P248`) two different subjects, the Metropolitan Museum of Art (`wd:Q160236`) and Christie's (`wd:Q503176`), and asserts (as a *situated statement*) that creator Joseph Vernet (`wd:Q315819`) is deprecated. The second *situating statement* is produced by an auction house, has a provenance (`wd:Q1991865`) and a date (`pr:P585`), and asserts as a *situated statement* that Charles Eschard (`wd:Q2959023`) is the favored creator. Fig. 1 below shows the visualization employed by GraphDB using Named Graphs:

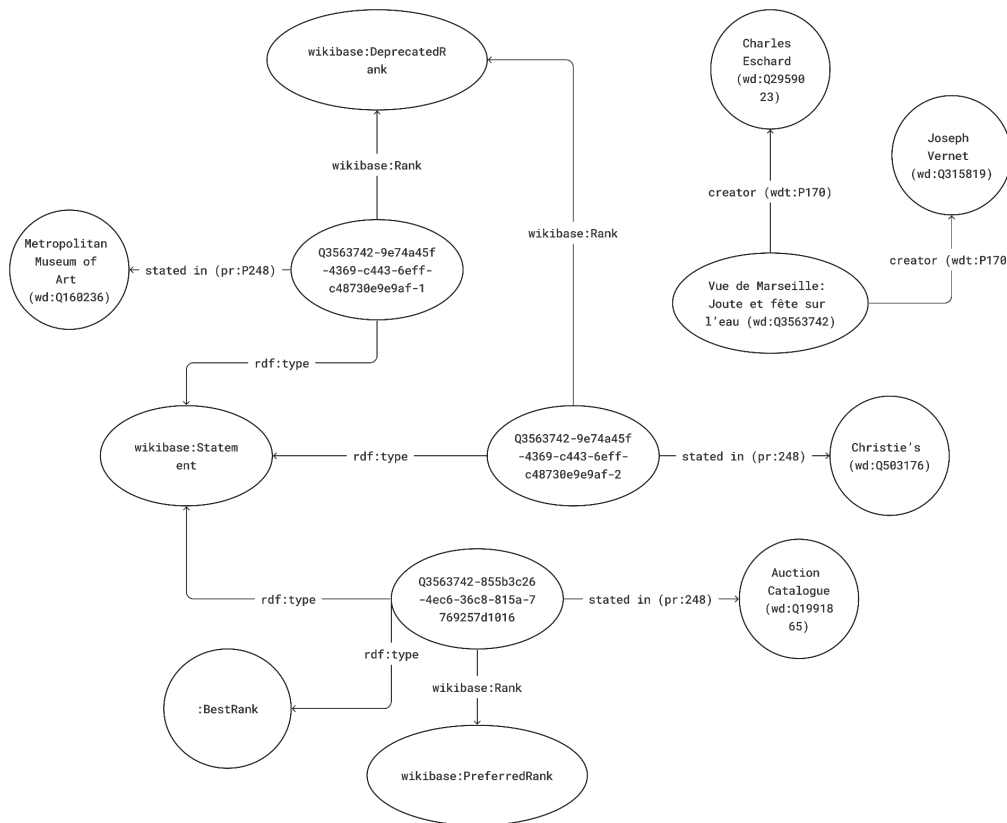


Figure 1. A diagram showing a visualization produced by GraphDB using Named Graph.

As shown, using Named Graphs, the situated statements (or triples regarding authors) are not related to the contested situating statements. By exploiting RDF-star instead, the visualisation shows the possibility of adding properties directly to the edges. Nevertheless, traditional graphs fail to address the situating statement by placing it at the same level as the situated statement. In particular, we claim that the situating statement can be conveyed much more effectively by exploiting *higraphs* [5]. Indeed, by enclosing the situated graph within a closed curve, it is possible to create sets that can have, in turn, edges to each other or other nodes. A higraph (also known as a hypergraph) is a generalization of a graph data structure which extends ordinary mathematical graphs by including the notions of depth and orthogonality. As such, nodes in a higraph are partitioned into different orthogonal planes. In other words, while a standard graph consists of nodes connected by edges, a hypergraph allows for connections between nodes and sets of nodes, referred to as hyperedges. Thus, nodes can be encapsulated into other nodes, thus defining a hierarchy. A higraph can therefore become a structure, bounded by its area, suitable for naming and labelling purposes [5].

As such, using categories for uncertainty visualisation defined by [9], we propose higraph as a visualisation technique to be *explicit* (i.e., uncertainty is explicitly signified through the enclosure of the situated statement), *extrinsic* (i.e., situating statements are extrinsically depicted by mapping them to new objects in the display) and *coincident* (i.e., situated and situating statements are visualised in the same space).

By employing a higraph as a visualisation technique, it is possible to visualise the first two assertions as shown in Fig. 2 below:

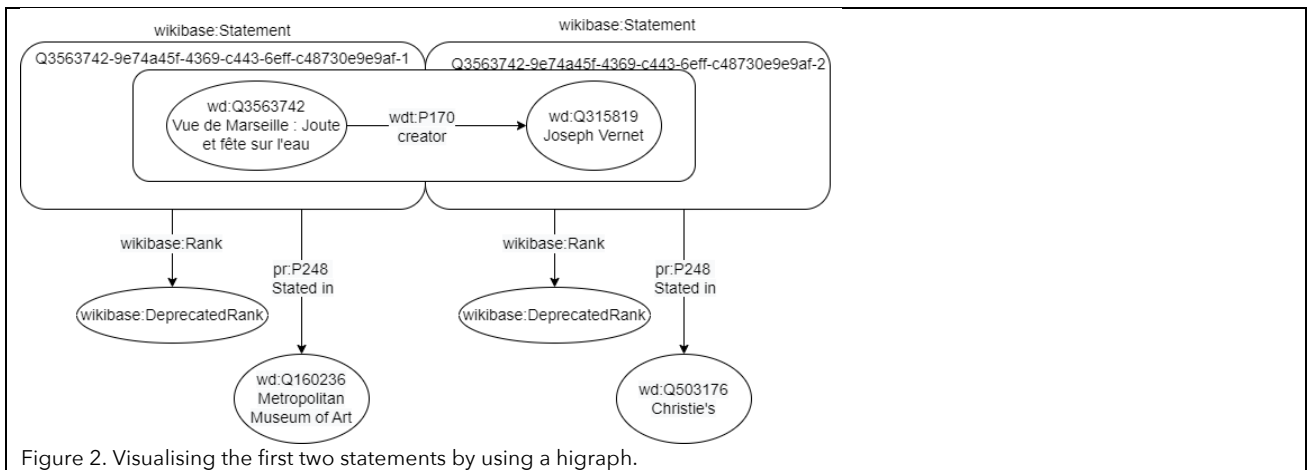


Figure 2. Visualising the first two statements by using a higraph.

Again through higraph visualisation, the situated statement about the preferred creator of the third assertion can be enclosed in the situating statement as shown below in Fig. 3:

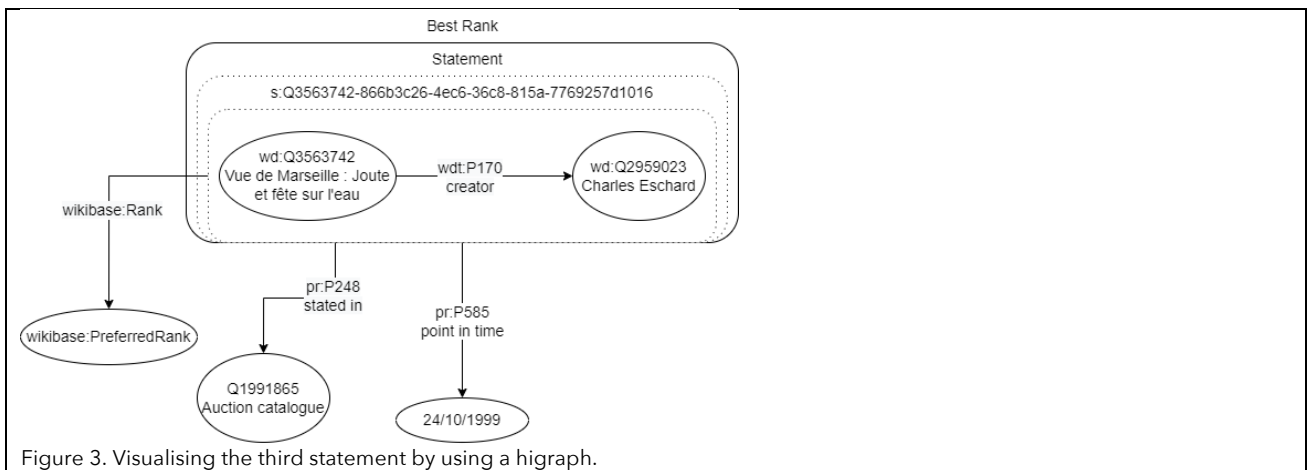


Figure 3. Visualising the third statement by using a higraph.

#### 4. CONCLUSIONS

Despite the spread of new paradigms for modelling qualitative data in RDF, there is still a lack of effective approaches and tools for visualizing complex, uncertain information. In this paper, we have shown a new non-representational approach to support the visualisation of qualitative information, preferring to build on top of existing large-scale strategies rather than opting for a disruptive approach. Higraphs provide an extension of existing visual metaphors and standards while offering a novel perspective to domain experts for exploring situated and situating statements. Thus, this graph formalism can be a particularly suitable solution to describe different approaches to meta-information about entities and relationships, efficiently conveying complex information through a scalable visualization technique that can be integrated with strategies already used on a large scale.

#### BIBLIOGRAPHY

- [1] Barabucci, Gioele, Francesca Tomasi, e Fabio Vitali. «Supporting Complexity and Conjectures in Cultural Heritage Descriptions», COLCO 2020. Collect and Connect: Archives and Collections in a Digital Age (2021): 104-115. <https://hdl.handle.net/11585/820304>
- [2] Daquino, Marilena, Valentina Pasqual, e Francesca Tomasi. «Knowledge Representation of Digital Hermeneutics of Archival and Literary Sources». J LIS, vol. 3 (2020). <https://doi.org/10.4403/jlis.it-12642>
- [3] Drucker, Johanna. «Graphical Approaches to the Digital Humanities». In A New Companion to Digital Humanities, edited by Susan Schreibman, Ray Siemens, e John Unsworth, 238–50. Chichester, UK: John Wiley & Sons, Ltd, 2015. <https://doi.org/10.1002/9781118680605.ch17>
- [4] Drucker, Johanna. «Non-Representational Approaches to Modeling Interpretation in a Graphical Environment». Digital Scholarship in the Humanities 33, 2 (2018): 248–63. <https://doi.org/10.1093/llc/fqx034>
- [5] Harel, David. «On Visual Formalisms». Communications of the ACM 31, 5 (1988): 514–30. <https://doi.org/10.1145/42411.42414>

- [6] Hassenzahl, Marc, e Noam Tractinsky. «User Experience - a Research Agenda». *Behaviour & Information Technology* 25, 2 (2006): 91–97. <https://doi.org/10.1080/01449290500330331>
- [7] Hullman, Jessica. «Why Authors Don't Visualize Uncertainty». *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020): 130–39. <https://doi.org/10.1109/TVCG.2019.2934287>
- [8] Kinkeldey, Christoph, Alan M. MacEachren, e Jochen Schiewe. «How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies». *The Cartographic Journal* 51, 4 (2014): 372–86. <https://doi.org/10.1179/1743277414Y.0000000099>
- [9] Martin-Rodilla, Patricia, e Cesar Gonzalez-Perez. «Representing Imprecise and Uncertain Knowledge in Digital Humanities: A Theoretical Framework and ConML Implementation with a Real Case Study». In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 863–71. Salamanca Spain: ACM, 2018. <https://doi.org/10.1145/3284179.3284318>
- [10] Po, Laura, Nikos Bikakis, Federico Desimoni, e George Papastefanatos. «Empirical Evaluation of Linked Data Visualization Tools». In *Linked Data Visualization*, di Laura Po, Nikos Bikakis, Federico Desimoni, e George Papastefanatos, 97–108. *Synthesis Lectures on Data, Semantics, and Knowledge*. Cham: Springer International Publishing, 2020. [https://doi.org/10.1007/978-3-031-79490-2\\_5](https://doi.org/10.1007/978-3-031-79490-2_5)
- [11] Sabburg, Letitia, Alan Woodley, e Kerrie Mengersen. «A Data- and Task- Oriented Design Framework for Bivariate Communication of Uncertainty», 2021. <https://doi.org/10.48550/ARXIV.2112.06921>
- [12] Tomasi, Francesca. «Archivi di persona in linked open data: il modello concettuale come strumento di integrazione nei GLAMs». *AIB studi* 57, 2 (2017). <https://doi.org/10.2426/aibstudi-11647>
- [13] Van Ruymbeke, M., P. Hallot, e R. Billen. «Enhancing CIDOC-CRM and Compatible Models with the Concept of Multiple Interpretation». *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences IV-2/W2* (2017): 287–94. <https://doi.org/10.5194/isprs-annals-IV-2-W2-287-2017>
- [14] Whitelaw, Mitchell. «Generous Interfaces for Digital Cultural Collections», *Digital Humanities Quarterly* 9, n. 1 (2015).
- [15] Windhager, Florian, Paolo Federico, Gunther Schreder, Katrin Glinka, Marian Dork, Silvia Miksch, e Eva Mayr. «Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges». *IEEE Transactions on Visualization and Computer Graphics* 25, 6 (2019): 2311–30. <https://doi.org/10.1109/TVCG.2018.2830759>
- [16] Windhager, Florian, Saminu Salisu, e Eva Mayr. «Exhibiting Uncertainty: Visualizing Data Quality Indicators for Cultural Collections». *Informatics* 6, 3 (2019): 29. <https://doi.org/10.3390/informatics6030029>

# **Soavissime ariette: impiego di repertori digitali per lo studio della tradizione delle arie d'opera**

Angelo Pompilio<sup>1</sup>, Giovanna Casali<sup>2</sup>, Paolo Bonora<sup>3</sup>

<sup>1</sup> Dipartimento di Beni Culturali, Università di Bologna, Italia – angelo.pompilio@unibo.it

<sup>2</sup> Dipartimento di Beni Culturali, Università di Bologna, Italia – giovanna.casali2@unibo.it

<sup>3</sup> Dipartimento di Filologia Classica ed Italianistica Università di Bologna, Italia – paolo.bonora@unibo.it

## **ABSTRACT**

Le arie assolvono una funzione strutturale nel melodramma: sono il luogo deputato all'espressione delle emozioni e degli stati d'animo dei personaggi. Questa specifica funzione, la loro natura lirica contrapposta ai recitativi, l'essere il momento di massimo impegno vocale per gli interpreti, sono gli elementi che contribuiscono a determinarne il particolare fenomeno del riuso. Il contributo presenta i risultati ottenuti nello studio del fenomeno nella prima metà del Settecento attraverso una repertoriazione estensiva delle arie dei libretti d'opera a stampa. Grazie allo spoglio sistematico, alla descrizione analitica dei contenuti e alla trascrizione del testo completo è stata allestita una base informativa che consente l'analisi delle relazioni tra le diverse varianti. Per tracciarne la tradizione è stata proposta una classificazione della natura degli interventi attuati dai poeti e un modello descrittivo in grado di rappresentarla in forma stemmatica. A partire da questa base dati è stato realizzato un sistema di consultazione per consentire il confronto sistematico delle opere complete o delle singole arie. Il risultato è uno strumento utile ad avviare indagini sistematiche per l'attribuzione delle responsabilità autoriali in grado di restituire plasticamente le dinamiche del fenomeno del riuso del testo poetico per il melodramma su una scala ad oggi mai sperimentata.

## **PAROLE CHIAVE**

Music drama; 18th century opera; opera arias; libretto; corago

## **1. INTRODUZIONE**

Il teatro d'opera è senza dubbio la forma di spettacolo più importante nella tradizione musicale occidentale ed incide in modo significativo nella vita sociale e culturale italiana ed europea dal Seicento in avanti. Inaugurato a Firenze nel 1600, esso si consolida ed espande in tutta Europa nel giro di pochi decenni e la sua tradizione prosegue senza interruzioni fino ad oggi grazie all'attività continuativa e a un efficace e complesso sistema di organizzazione e produzione [2]. Il teatro d'opera combina dramma, musica e rappresentazione visiva, tre elementi capaci di sedurre lo spettatore per l'intricata e avvincente storia di due innamorati che agiscono nella cornice di una vicenda storica, eroica o mitologica dell'antichità o di epoca moderna, per le suadenti melodie e i mirabolanti virtuosismi vocali dei cantanti nonché per la magnificenza degli apparati scenici.

Alla base dell'opera in musica c'è un dramma, un testo letterario (il 'libretto'), interamente cantato, dall'inizio alla fine. Nella sua lunga parabola storica il dramma musicale ha impiegato dispositivi musicali e drammaturgici diversi ma capaci di rappresentare in modo efficace lo sviluppo dell'azione attraverso dialoghi serrati tra i personaggi e momenti più lirici, caratterizzati da una sospensione dell'azione che consente ai personaggi di esprimere il proprio stato d'animo, le emozioni, le passioni di quel particolare momento del dramma [1][4].

Nei primi anni del Settecento e fino ai primi decenni del secolo successivo la produzione operistica si divarica in due generi autonomi e distinti sotto il profilo formale e organizzativo: l'opera seria e l'opera comica [12]. Da un lato le storie esemplari ed edificanti degli eroi dell'antichità affidate alla vocalità florida e funambolica delle prime donne e dei cantanti evirati che agiscono nella cornice di apparati scenografici sontuosi dei grandi teatri, dall'altra le vicende minute della vita quotidiana di servitori, fornai, barbieri e dottori, che si incarnano nella spigliatezza e nel brio dei cantanti comici. Due i modelli di riferimento nella produzione librettistica di questi due generi: i drammi per musica di Pietro Metastasio per l'opera seria, i drammi giocosi di Carlo Goldoni per l'opera comica.

L'indagine che qui viene presentata ha per oggetto l'opera seria della prima metà del Settecento. L'attività svolta finora è stata condotta in particolare sulle fonti librettistiche degli anni compresi tra il 1715 e il 1738 [5][8][9].

Il dispositivo musicale e drammaturgico dell'opera seria è molto semplice e lineare. Il dramma è articolato in tre atti, ciascun atto a sua volta si articola all'incirca in una quindicina di scene e ciascuna scena di solito si apre con un recitativo e si chiude con un'aria. Il recitativo è in versi sciolti, settenari ed endecasillabi, l'aria è un testo strofico, di solito due strofe



di tre-cinque versi isometrici, in Metastasio prevalentemente di metro settenario, ma anche quinari, senari, ottonari e decasillabi. Al recitativo è affidato lo svolgimento dell'azione attraverso il dialogo tra i personaggi, l'aria invece esprime lo stato d'animo, le emozioni, l'affetto di un singolo personaggio in quel particolare momento. Musicalmente il recitativo consiste nell'intonazione continuativa del testo in uno stile vocale a metà strada tra il parlato e il cantato, l'aria invece è un brano di forma chiusa per una sola voce con orchestra, che gode di una piena autonomia e unitarietà nella partitura musicale, ed ha il compito di esprimere, attraverso la musica, il contenuto affettivo del testo [7]. Tutte le potenzialità della musica di suscitare emozioni, suggestioni e incanto sono dispiegate al meglio nelle arie attraverso profili melodici, andamenti agogici e dinamici, soluzioni timbriche e successioni armoniche capaci di catturare l'attenzione di chi ascolta. La magia della fascinazione che le "soavissime ariette"<sup>1</sup> sono capaci di generare nello spettatore si realizza però in teatro solo grazie alle capacità interpretative del cantante e al virtuosismo canoro che dispiega.

Questi aspetti drammaturgico-musicali si intersecano però con aspetti di ordine pratico e convenzioni teatrali consolidate. Nel cast di una rappresentazione teatrale ci sono prime parti e parti secondarie. Alle prime parti compete un ruolo di rilievo con l'intonazione di almeno cinque-sei arie di dimensioni più vaste e con caratteristiche tecniche di particolare impegno, alle parti secondarie un numero di arie inferiore e di difficoltà più contenuta. Mediamente i personaggi di un'opera sono sei, di conseguenza un dramma per musica si presenta come successione di una trentina di arie intervallate da recitativi più o meno estesi. Nella stesura del dramma il poeta deve predisporre le arie in un ordine tale che il contenuto affettivo risulti sempre diverso, e magari contrastante, con l'aria che precede e segue [6] [11] [13]. Di conseguenza, il testo del recitativo che precede l'aria deve determinare la situazione affettiva che il personaggio esprimerà. Nella disposizione delle arie bisognerà fare attenzione anche all'avvicendamento dei personaggi, sia per dare varietà all'azione, sia per consentire al cantante di riguadagnare le energie impiegate nello sforzo fisico sostenuto.

A quest'epoca il testo di un dramma per musica viene scritto in un arco temporale molto stretto, tre-quattro settimane, a causa dei ritmi serratissimi dettati dall'organizzazione del teatro impresariale. Il libretto messo in musica può proporre soggetti nuovi, ricavati liberamente da soggetti storici dell'antichità o drammi di epoche più recenti, rimaneggiare soggetti ripresi o rielaborati da libretti anteriori di alcuni anni o lustri, o semplicemente riproporre, con interventi di entità contenuta, un libretto preesistente con musica completamente nuova. Quest'ultimo caso risulta molto frequente nella produzione operistica del Settecento. Soltanto un esempio per fornire un'indicazione quantitativa del fenomeno: i 27 drammi per musica scritti da Pietro Metastasio hanno dato luogo a più di mille opere liriche di autori diversi.

Nel reimpiego di tale materiale, il poeta o il musicista rimaneggiano il testo originario in maniera assolutamente libera, sia dal punto di vista dell'azione drammatica sia rispetto alle arie riproposte o di nuovo conio. Come s'è detto, le arie sono organismi autonomi all'interno del dramma in quanto esprimono un contenuto affettivo definito in una forma musicale chiusa. Tale peculiarità consente il riuso del testo e della musica di un'aria, o soltanto di uno dei due, in drammi diversi dal dramma originario nel quale l'aria è stata impiegata per la prima volta. L'aria va intesa, pertanto, come un elemento funzionale ad un determinato contesto drammatico. Arie diverse, che esprimono il medesimo contenuto affettivo, possono essere impiegate liberamente dal librettista o dal compositore in libretti diversi là dove il contesto drammatico del libretto risulti coerente con l'aria prescelta: le arie sono 'fungibili' a parità di contenuto.

Nei libretti consultati in questa prima fase dell'indagine sono stati riscontrati molti casi diversi di sostituzione e reimpiego di arie. Per dar conto dei fenomeni osservati sono state definite alcune categorie di intervento, fra le quali: variante, variante significativa, riscrittura (il testo presenta numerose varianti, con termini sostituiti e/o posti in altra posizione) o semplice affinità (testi in cui restano solo lacerti o flebili reminiscenze dell'antecedente).

## 2. PER UNA TRADIZIONE TESTUALE DELL'ARIA

Per ricostruire le dinamiche di riuso del testo delle arie è necessario realizzare una repertoriazione estensiva, la più completa possibile, in grado di fornire un quadro attendibile del fenomeno. Il progetto si è posto dunque l'obiettivo di repertoriare questa produzione nella prima metà del Settecento, al momento dal 1715 al 1738, attraverso l'indagine sistematica condotta sui libretti coevi. Partendo da un corpus di 5200 libretti (di cui circa 4000 disponibili in riproduzione digitale), di cui 840 sono prime edizioni assolute, 1500 circa le nuove intonazioni di libretti già musicati e 2860 le edizioni successive. Da questo corpus si potranno ricavare all'incirca 160.000 arie (dai soli libretti riprodotti 120.000), stima ottenuta considerando una media per difetto di 30 arie per opera. Ad oggi sono state inserite nel sistema 62.025 arie da 1880 libretti. Di queste possono essere considerate arie nuove 29.991, poco meno del 50% del totale, 18.912 sono derivate da esse con o senza varianti, le rimanenti 13.122 sono le arie riproposte in successive rappresentazioni. Questi numeri danno la dimensione del fenomeno e ne motivano la rilevanza nello studio della funzione dell'aria nel melodramma settecentesco, non solo dal

---

<sup>1</sup> Così Lorenzo Da Ponte apostrofa le arie del Metastasio in contrasto con la produzione coeva non ritenuta all'altezza dell'archetipo (Da Ponte, Lorenzo. *Memorie*. Gambarin, Giovanni e Nicolini, Fausto (a cura di). vol. I. Bari: G. Laterza. 1918. p. 79).

punto di vista del valore del testo poetico in relazione al testo musicale, ma anche in prospettiva storica e filologica alla luce delle prassi produttive e dei rapporti di forza tra i diversi attori coinvolti. Per studiare questa complessa rete di relazioni su un numero così significativo di testi è quindi indispensabile impiegare uno strumento informatico. Partendo dalla repertoriazione analitica delle fonti testuali già presenti nell'archivio Corago<sup>2</sup> [3], si sono inseriti i testi completi delle arie e riscontrate le derivazioni classificandole in base alle categorie già menzionate. Al momento, questa attività è stata condotta attraverso una analisi critica supportata da uno strumento di confronto delle corrispondenze verso per verso. L'analisi è, inoltre, supportata dalla relazione con l'interprete derivata dallo spoglio del corrispondente libretto che, di norma, riporta l'elenco dei personaggi e dei relativi interpreti. Lo studio della tradizione di questi testi può essere inoltre contestualizzato nel processo storico di impiego, ovvero la loro messa in scena. La ricostruzione filologica può quindi avvalersi del dato storico e consentire di indagare fenomeni come le cosiddette "arie di baule", ovvero quelle arie che venivano eseguite in opere ad esse estranee ma cavalli di battaglia dei singoli interpreti.

D'altra parte, la paternità musicale della singola aria va verificata attraverso la consultazione della partitura musicale, non sempre disponibile. Ad oggi, possiamo solo indicare che una stessa aria è stata cantata dal medesimo interprete in opere musicate da autori diversi. Nell'analisi condotta finora, l'informazione è stata ricavata dal solo libretto e le attribuzioni espresse dovranno pertanto essere verificate puntualmente sulle fonti musicali, se pervenute, o su altre fonti coeve che ne testimoniano la paternità.

### 3. STRUMENTI E RISORSE

Per ricostruire la fitta trama di dipendenze tra i testi delle arie sono stati sviluppati un modello concettuale e degli strumenti che consentono di analizzare la loro forma e i contenuti alla luce del rispettivo contesto esecutivo<sup>3</sup>. In questo modo, l'analisi morfologica e semantica del testo trova supporto e riscontro nell'informazione storica della cronologia degli spettacoli derivata dalla catalogazione analitica dei rispettivi libretti. Al modello originario dell'archivio Corago, articolato nelle tre sezioni, Repertorio, Spettacoli e Libretti, è stata aggiunta un'area trasversale che estende da un lato la descrizione del libretto introducendo lo spoglio analitico di tutte le arie presenti, dall'altro rinvia allo spettacolo attraverso la relazione con l'interprete e al repertorio attraverso il personaggio interpretato. Entrambe le relazioni rinviano a loro volta all'authority file dei nomi in cui confluiscono tutte le responsabilità e i relativi ruoli attestati nelle fonti. Infine, una nuova sezione è dedicata alla rappresentazione delle relazioni individuate tra le arie. A supporto della ricerca delle relazioni è stata sviluppata una semplice procedura di confronto tra i testi a livello di singolo verso che ne intercetta la distanza tra le rispettive stringhe. Per la ricerca e navigazione delle informazioni delle nuove sezioni del modello è stato allestito un prototipo di applicazione web che prevede diverse chiavi di accesso: per aria (incipit, testo, metro), per opera (titolo, luogo e data di rappresentazione) o per autori e interpreti (nomi, personaggio interpretato). Il sistema consente di ricercare direttamente le arie o le opere che le contengono. In questo modo l'utente può avviare la sua ricerca partendo dal contenuto (la singola aria) o dal suo contenitore (l'opera). Le pagine relative alle schede delle due tipologie di entità riportano in modo sinottico tutti i metadati descrittivi e le relazioni presenti nel modello. Per la singola aria sono presenti: i dati fondamentali e il testo completo; i dati dell'opera, dello spettacolo e del libretto oltre ad una sezione che riproduce lo stemma dell'aria. In questo vengono rappresentate in forma gerarchica la collocazione della singola aria nell'albero delle relazioni coi predecessori e successori (Figura 1).

---

<sup>2</sup> Corago è il repertorio e archivio dei libretti del melodramma italiano dal 1600 al 1900. Accessibile online all'indirizzo: <http://corago.unibo.it/>. È disponibile anche una versione Linked Open Data all'indirizzo: <http://corago.unibo.it/lod>. I dati sono pubblicati in forma di dataset RDF reperibile attraverso il DOI: <https://doi.org/10.5281/zenodo.3865867>.

<sup>3</sup> Il prototipo di applicazione è disponibile all'indirizzo: <http://www.ilcorago.org/WPcorago/arie.asp>

Incipit	relazione	Personaggi	Interpreti	Autori	Opera Autori	Luogo Edificio	Data
<input type="checkbox"/> L1 Vanne alla pena, o perfido	nuova	Anassandro Trasimede	P. Casati	F. Gasparini A. Zeno	Merope** F. Gasparini [A. Zeno]	Venezia Teatro Tron di San Cassano	27/08/1719
=	riposta	=	C. A. Mazza	=	Merope	Firenze Teatro del Cocomero	26/02/1713
<input type="checkbox"/> L2 Perfido, è ver, cadrò	varianti significative	Anassandro	[non indicato]	Anonimo A. Zeno	Merope Anonimo [A. Zeno]	Milano Regio Ducale Teatro	06/01/1715
=	riposta	=	[non indicato]	=	Merope	Karlsruhe Teatro	1729
<input type="checkbox"/> L3 Perfido, è ver, cadrò	ripresa	Anassandro	I. Germani	Anonimo A. Zeno	Merope Anonimo A. Zeno	Napoli gran sala del Reggio palazzo	01/10/1716
<input type="checkbox"/> L3 Perfido, è ver, cadrò	ripresa	Anassandro	G. M. Boschi	G. M. Orlandini A. Zeno	La Merope* G. M. Orlandini A. Zeno	Bologna Teatro Formagliari	24/10/1717
=	riposta	=	A. Imperatori	=	Merope	Recanati Teatro	01/01/1719
<input type="checkbox"/> L3 Perfido, è ver, cadrò	ripresa	Anassandro	A. M. Mangani	L. A. Predieri A. Zeno P. Pariati	Merope* L. A. Predieri A. Zeno P. Pariati	Livorno Teatro San Sebastiano	1718
<input type="checkbox"/> L3 Cadrà quest'alma, è ver	riscrittura	Gerilda	F. Cuzzoni	G. Cozzi A. Zeno P. Pariati	Ambleto* G. Vignati C. Baliani G. Cozzi A. Zeno P. Pariati	Milano Regio Ducale Teatro	27/08/1719
<input type="checkbox"/> L4 Cadrà quest'alma, è ver	varianti	Anassandro	F. Cignani	P. Torri A. Zeno	La Merope* P. Torri A. Zeno	Monaco di Baviera Hof	12/10/1719
=	riposta	=	G. F. Costanzi	=	La Merope	Monaco di Baviera Hof	24/01/1723
<input type="checkbox"/> L5 Cadrà mia salma, è ver	varianti	Baiazet	G. Païta	F. Gasparini A. Piovene	Bajazette [F. Gasparini] [A. Piovene]	Venezia Teatro Grimani di S. Samuele	05/05/1723
<input type="checkbox"/> L4 Cadrà quest'alma, è ver	varianti significative	Anassandro	[non indicato]	Anonimo A. Zeno P. Pariati	Merope Anonimo A. Zeno	Firenze Teatro della Pergola	267 dic. 1729
<input type="checkbox"/> L3 Perfido, è ver, cadrò	ripresa	Anassandro	[non indicato]	Anonimo A. Zeno	Merope Anonimo A. Zeno	Cremsier [= Kromeritz]	estate 1727

Figura 1. Rappresentazione dello stemma di un'aria.

Per le opere, oltre ai dati salienti relativi agli autori, alle rappresentazioni e alla fonte, ai personaggi e agli interpreti, il sistema presenta la successione di tutte le arie suddivise per atti. Ciascuna entrata dell'elenco permette di accedere alla relativa scheda con lo stemma. In questo modo l'analisi può partire dall'opera attribuita ad un poeta o ad un compositore e verificare la collocazione delle singole arie nello stemma consentendo così di approfondire l'attribuzione di ciascuna. In modo simile è possibile percorrere il procedimento inverso, considerare la versione capostipite di un'aria e percorrere le attestazioni nelle opere in cui compare con la relazione indicata.

Alla base del processo di identificazione della relazione tra le diverse varianti c'è la funzionalità di comparazione dei testi. In questo caso l'interfaccia utente (Figura 2) consente di verificare a colpo d'occhio gli interventi occorsi alla luce dei dati dello spettacolo e dell'opera. Parallelamente, per le opere è possibile individuare la distribuzione delle singole arie nelle opere che condividono con essa almeno un'aria. Selezionando dall'elenco le opere di interesse, il sistema consente di comparare la rispettiva articolazione e identificare eventuali sostituzioni, variazioni o aggiunte nella sequenza delle arie.

## Compara Arie

<p>09/01/1712 – Venezia <b>F. Gasparini (comp.)</b> [A. Zeno] (lib.)</p> <p>Vanne alla pena, o perfido. Perfido, è ver, cadrò: non cadrò solo.</p> <p>Nel mio cader trarrò qualche piacer almen dall'altrui duolo.</p> <p>---</p>	<p>06/01/1715 – Milano <b>Anonimo (comp.)</b> [A. Zeno] (lib.)</p> <p>Perfido, è ver, cadrò ma nel mio fier destin non cadrò solo.</p> <p>Nel mio cader avrò qualche piacer alfin de l'altrui duolo.</p> <p>varianti significative </p> <p>nella 1a strofa, v. 2=1 Vanne alla pena, o perfido <b>F. Gasparini (comp.)</b>, <b>Merope**</b> 09/01/1712, Venezia</p>	<p>27/08/1719 – Milano <b>G. Cozzi (comp.)</b> A. Zeno (lib.) P. Pariati (lib.)</p> <p>Cadrà quest'alma, è ver, ma nel mio fier destin sola non perirà.</p> <p>De l'altrui duol alfin forse qualche piacer almen cadendo avrò.</p> <p>riscrittura </p> <p>Perfido, è ver, cadrò <b>Anonimo (comp.)</b>, <b>Merope</b> 06/01/1715, Milano</p>
---	--	--

Figura 2. Strumento di comparazione delle arie.

Nel complesso, sia il corpus allestito, che gli strumenti implementati, seppur in forma prototipale, forniscono agli studiosi materiali di prima mano e risorse che consentono di indagare il fenomeno in una prospettiva e su una scala ad oggi mai sperimentate.

#### 4. ESITI ED EVOLUZIONI

Ad oggi, il risultato ottenuto è l'allestimento di una base informativa intesa come strumento repertoriale per l'avvio di indagini che potranno estendersi considerando ulteriori fonti. Infatti, avendo utilizzato i libretti come fonte prevalente, le informazioni riportate possono risultare inesatte. Ad esempio, spesso i libretti che indicano Metastasio come autore del testo poetico contengono in realtà arie modificate in modo tanto significativo da invalidarne l'attribuzione o semplicemente sostituite con altre del tutto nuove o ricavate da libretti di altri autori. Solo analizzando gli stemmi è possibile valutare criticamente le attribuzioni delle singole arie. Lo strumento predisposto è quindi un valido punto di partenza per lo studio delle attribuzioni sia dei testi sia della musica delle arie.

Da una prima osservazione, il fenomeno del riuso della singola aria sembra manifestarsi in un arco temporale della durata di circa quarant'anni. Per consolidare la lettura di questa dinamica, i dati finora raccolti potranno giovare di una estensione cronologica dell'arco temporale considerato. Per raggiungere quest'obiettivo si prevede di estendere l'indagine all'intera produzione di libretti della prima metà del Settecento procedendo allo spoglio completo di almeno quelli già disponibili in riproduzione digitale.

Una volta arricchita la base informativa relativa alla tradizione testuale si potrà rivolgere lo sguardo alle fonti musicali per approfondire i casi di attribuzione incerta non solo della musica, ma anche del testo. L'obiettivo è aggiungere una ulteriore dimensione di analisi derivata da un testo di natura diversa, a sua volta collocato in una tradizione sua propria. I possibili esiti di una analisi comparata delle due tradizioni, portata a termine su questa scala, restano però tutti da approfondire.

La dimensione delle indagini prospettate, sia in termini di fonti che di affinamento dei metodi, richiederà l'impiego di risorse che potranno dispiegarsi soltanto con il contributo della più ampia comunità degli studiosi interessati.

#### BIBLIOGRAFIA

- [1] Bianconi, Lorenzo. *La drammaturgia musicale*. Bologna: Il Mulino, 1986.
- [2] Bianconi, Lorenzo e Walker, Thomas R. "Production, Consumption and political Function of Seventeenth Century Opera", *Early Music History*, IV, 1984, pp. 209-96.
- [3] Bonora, Paolo e Pompilio, Angelo. "Corago in LOD. The debut of an Opera repository into the Linked Data arena", *JLIS.it*, 12, 2 (May 2021): 54-72. DOI: 10.4403/jlis.it-12699
- [4] Dahlhaus, Carl. *Drammaturgia dell'opera italiana*, in Bianconi, Lorenzo e Pestelli, Giorgio (a cura di) *Storia dell'opera italiana*, vol. VI. Torino: EDT, 1988, pp. 77-162.
- [5] Della Seta, Fabrizio. *Il librettista*, in Bianconi, Lorenzo e Pestelli, Giorgio (a cura di), *Storia dell'opera italiana*, vol. IV. Torino: EDT, 1987, pp. 233-91.
- [6] Di Benedetto, Renato. "Dal Metastasio a Pergolesi e ritorno. Divagazioni fra l'Adriano in Siria e L'olimpiade", *Il Saggiatore musicale*, II, 1995, pp. 259-295.
- [7] Di Benedetto, Renato. "Il Settecento e l'Ottocento", in *Letteratura italiana*, vol. 6, Torino, Einaudi, 1986, pp. 365-410.
- [8] Fabbri, Paolo. *Il secolo cantante: per una storia del libretto d'opera nel Seicento*. Bologna: Il Mulino, 1990.
- [9] Gallarati, Paolo. *Musica e maschera. Il libretto italiano del Settecento*. Torino: EDT Musica, 1984.
- [10] Goldin, Daniela. *Per una morfologia dell'aria metastasiana*, in Muraro, Maria Teresa (a cura di), *Metastasio e il mondo musicale*. Firenze: Leo Olschki, 1986, pp. 13-37.
- [11] Strohm, Reinhard. "Aria e recitativo. Dalle origini all'Ottocento", in Nattiez, Jean-Jacques (a cura di), *Enciclopedia della musica*, Torino, Einaudi, 2004, vol. IV: Storia della musica europea, pp. 416-429.
- [12] Strohm, Reinhard. *Dramma per musica: Italian opera seria of the eighteenth century*. New Haven (etc.): Yale University Press, 1997.
- [13] Strohm, Reinhard. *Italienische Opernarien des frühen Settecento: 1720-1730*. Köln: Arno Volk Verlag Hans Gerig, 1976.

# *Musisque Deoque:* la nuova organizzazione dei testimoni manoscritti

Alessia Prontera<sup>1</sup>, Valeria Melis<sup>2</sup>

<sup>1</sup> Università Ca' Foscari Venezia, Italia – alessia.prontera@unive.it

<sup>2</sup> Università Ca' Foscari Venezia/ Università di Cagliari, Italia – valeria.melis@unive.it

## ABSTRACT\*

Il poster illustra i recenti lavori di aggiornamento e riorganizzazione dei metadati relativi ai manoscritti che trasmettono i testi contenuti nella risorsa *Musisque Deoque*.

## PAROLE CHIAVE

Archivi digitali; metadattazione; poesia latina; manoscritti

## 1. LO STATUS QUAESTIONIS

Con il presente lavoro si vogliono illustrare i recenti lavori di aggiornamento e di riorganizzazione dei manoscritti contenuti negli apparati dei testi contenuti in *Musisque Deoque*, un archivio digitale di poesia latina dalle origini al Rinascimento italiano.<sup>1</sup> La revisione prosegue e completa gli interventi operati, circa un decennio fa, da Linda Spinazzé,<sup>2</sup> curatrice dell'edizione digitale *MQDQ* delle *Elegie* di Massimiano (<https://www.mqdq.it/texts/MAXIM|eleg|001>). Attorno al 2012 il numero di testimoni era nettamente inferiore rispetto a quello attuale e superava di poco le 130 unità. Grazie alla sempre più consistente acquisizione del numero di apparati, la cifra dei testimoni è incrementata esponenzialmente, attestandosi attorno alle 4840 unità. Giova precisare che per testimone si intende *in primis* un singolo manoscritto (membranaceo o cartaceo), identificabile con precisione e ancora consultabile; in rari casi per testimone si può intendere anche un papiro: è il caso, *e.g.*, del Pap. Berol. 21138A-B conservato al Ägyptisches Museum und Papyrussammlung di Berlino, presente tra i testimoni dell'*Eneide* di Virgilio, curata dai proff. Massimo Gioseffi, Isabella Canetta e dalla dott.ssa Silvia Arrigoni. In occasione della nuova ristrutturazione, Luigi Tassarolo, tecnico informatico di *MQDQ*, ha suddiviso la vasta categoria dei testimoni in tre tipologie: testimoni strutturati, testimoni semi-strutturati e testimoni non strutturati. Il testimone strutturato è una fonte scomposta in un numero progettato di componenti obbligati, il cui valore è scelto entro una limitata gamma di valori predefiniti, che nel nostro caso sono voci di tabelle condivise da tutti gli oggetti dello stesso tipo. Il testimone semi-strutturato è mancante di almeno un valore. Entrambe le tipologie sono conservate nell'archivio delle fonti, dal quale attingono tutti gli apparati dei testi. Nella terza tipologia, che non rientra nell'archivio, rientrano in genere i testimoni non più identificabili, andati persi o distrutti; tuttavia, tra i testimoni non strutturato compaiono anche le *secundae manus* e i *correctores*, cioè gli interventi dei copisti medievali o rinascimentali apportati su un ms. registrato come testimone strutturato. Restando sull'apparato dell'*Eneide*, la differenza appare chiara con il *Vaticanus lat. 3225 (F)* e i *correctores* di epoca tardoantica (**F**<sup>1</sup>, **F**<sup>2</sup>, **F**<sup>3</sup>), medievale (**F**<sup>4</sup>) e rinascimentale (**F**<sup>5</sup>). In questa sede ci si concentrerà sulla tipologia dei testimoni strutturati.

**F** = Vaticanus lat. 3225 Romae (apud Symmachum?) exaratus [CLA I 11]

**F**<sup>1</sup> = correctores Vaticani lat. 3225 saec. IV-VI

**F**<sup>2</sup> = correctores Vaticani lat. 3225 saec. IV-VI

**F**<sup>3</sup> = correctores saec. Vaticani lat. 3225 VII-VIII

**F**<sup>4</sup> = additamenta Aevi Carolini

**F**<sup>5</sup> = correctores Vaticani lat. 3225 Renatarum Litterarum Aetatis

Figura 1. Il testimone **F** = *Vaticanus lat. 3225* nelle fonti dell'*Eneide*.

## 2. METADAZIONE: IL BACK-END

Nel back-end, nel quale si accede dopo aver inserito l'account e la password fornite dall'amministratore (<https://www.mqdqgalaxy.it/public/backend>), un testimone si struttura seguendo una maschera composta da una serie di metadati indispensabili: città (*Settlement*), biblioteca (*Repository*), fondo (*Collection*), segnatura (*Shelfmark*), datazione (*Date*), nome (*Name*), [*perma*]link (*URL*), illustrati dalla fig. 2:

\* Ogni parte di questo contributo è frutto della collaborazione tra le due autrici. Tuttavia, i paragrafi 1 e 2 sono stati curati nello specifico da Alessia Prontera, i paragrafi 3 e 4 da Valeria Melis.

<sup>1</sup> Sull'archivio *MQDQ* vd. almeno Mastandrea 2015 e Boschetti – Del Grosso – Spinazzé 2021.

<sup>2</sup> Spinazzé 2011.

Figura 2. La maschera per l’inserimento dei metadati dei manoscritti in *Musique Deoque*.

Va precisato che nel decennio intercorso tra l’inizio dei lavori e lo stato attuale, molte mani si sono avvicinate nella Galassia *MQDQ* inserendo nuovi apparati e testimoni senza seguire linee guida generali e, dunque, classificando in modo eterogeneo i testimoni e le edizioni critiche. La recente riorganizzazione si è posta in primo luogo l’obiettivo di omogeneizzare e regolarizzare la classificazione dei metadati indispensabili per l’identificazione del testimone. Rispetto ai lavori iniziati più di un decennio fa, ai metadati fondamentali, corrispondenti alle prime quattro voci elencate *supra*, sono state aggiunte tre voci: la datazione, la dicitura latina e il link alla digitalizzazione del manoscritto.

Il primo campo, il *Settlement*, è indicato sempre dal nome della città nella lingua nazionale (ad es. London, Montpellier, Wolfenbüttel ecc.); il *Repository*, la biblioteca presso la quale è conservato il testimone, è indicato con il nome moderno, in casi sporadici affiancato da quello precedente. Il campo *Collection* è presente solo laddove la organizzazione biblioteca preveda la presenza di fondi (es. Ff. alla University Library di Cambridge e Rep. alla Universitätsbibliothek di Lipsia); nel campo della segnatura, elemento dirimente, viene indicata sempre quella moderna, affiancata eccezionalmente da quella antica laddove indicata nell’archivio digitale della singola biblioteca (es. XII,10 (=3972) nella Biblioteca Nazionale Marciana di Venezia).

Tra i campi nuovi compare la datazione del manoscritto. Precedentemente all’intervento di riorganizzazione l’informazione compariva solo a discrezione del curatore dell’edizione digitale e, in genere, era tratta dall’edizione di riferimento. Per i manoscritti che ne erano privi si è deciso di integrare l’informazione reperendo la datazione dall’edizione cartacea e, laddove non indicata, si è inserita quella indicata nella scheda del manoscritto nel website ufficiale della biblioteca. Anche la data viene indicata secondo una rigida “grammatica” che prevede l’indicazione del secolo in numeri romani, preceduti da *saec.* e seguiti da *in.* (*ineunte*), *ex.* (*exeunte*), e *me.* (*medio*) nel caso in cui si possa specificare maggiormente la cronologia all’interno del secolo. Nei casi fortunati in cui sia noto l’anno di composizione, esso compare in numeri arabi, eventualmente preceduti da *post* o *ante*, qualora la datazione si aggiri attorno ad un anno.

Nel campo *Name* compare solitamente la denominazione in latino del manoscritto (*Laurentianus*, *Escorialensis*, *Antverpiensis* ecc.) seguito dalla segnatura. Solitamente in questa stringa vengono mantenute le informazioni inserite dal primo curatore dell’edizione digitale, che si è scelto di mantenere pressoché inalterate per evitare la soppressione del lavoro precedente.

L’ultimo campo è riservato al *permalink*, ovvero al collegamento digitale permanente, alla pagina del manoscritto contenuta nel sito ufficiale delle maggiori biblioteche europee, nord-americane (Boston, New York) e africane (Al-Qahira, Cape Town). Laddove possibile, si è preferito inserire il *link* alla pagina del manoscritto piuttosto che alla digitalizzazione affinché, in questo modo, l’utente di *MQDQ* prima di tutto possa reperire informazioni utili alla conoscenza filologiche paleografiche da lì accedere alla digitalizzazione del testimone, qualora sia messa a disposizione dalla biblioteca.

Vale la pena riportare alcuni tra i siti ufficiali delle biblioteche che sono stati impiegati per i *permalinks* dei manoscritti: anche numericamente prevale l’impiego di *permalinks* dal sito della Biblioteca Vaticana (<https://digi.vatlib.it/>); seguono su suolo nazionale la Biblioteca Medicea Laurenziana (<https://www.bmlonline.it/la-biblioteca/cataloghi/>) e la Veneranda Biblioteca Ambrosiana (<https://ambrosiana.comperio.it/biblioteca-digitale/>). In area germanofona spiccano la Bayerische Staatsbibliothek a Monaco (<https://opacplus.bsb-muenchen.de/metaopac/start.do>) e l’Universitätsbibliothek di Heidelberg (<https://www.ub.uni-heidelberg.de/helios/digi/handschriften.html>). In territorio francese è stato indispensabile l’impiego del sito della Bibliothèque National de France di Parigi (<https://archivesetmanuscrits.bnf.fr/>); per la penisola iberica si ricorda la Real Biblioteca del Monasterio de San Lorenzo de El Escorial (<https://rbmecat.patrimonionacional.es/>). Un

consistente apporto deriva anche oltremarina dalla British Library di Londra ([https://searcharchives.bl.uk/primo\\_library/libweb/action/search/.do](https://searcharchives.bl.uk/primo_library/libweb/action/search/.do)). Nei casi non sia stato possibile reperire né la digitalizzazione né la scheda del testimone, si sono rivelati utili il ricco catalogo messo a disposizione da *Mirabile. Archivio digitale della cultura medievale* (<http://www.mirabileweb.it/index.aspx>) e il corposo portale di *Biblissima* (<https://portail.biblissima.fr/>).

### 3. METADAZIONE: IL FRONT-END

Nel *front-end* delle pagine web di *MQDQ* riportanti i testi e gli apparati di ciascuna opera, la visualizzazione delle informazioni sui manoscritti può avvenire cliccando sul tasto *Testimoni* (in alto a destra) e poi selezionando dal menù il testimone d'interesse oppure direttamente dagli apparati al testo. Concentrandoci, per adesso, su quest'ultimo caso, prendiamo a mo' di esempio l'apparato relativo a Lucan. 1,320<sup>3</sup> e supponiamo che l'utente voglia visualizzare le informazioni relative al manoscritto P. Al passaggio del cursore sul *siglum*, compaiono due diciture, la prima corrispondente al nome del manoscritto completo di tutte le informazioni (in rosso), la seconda, corrispondente al nome latino (in nero), che offre i metadati relativi alla fonte *precedentemente* l'intervento di riorganizzazione (fig. 3):

The screenshot shows the MQDQ interface with a Latin text passage. The text is: "Quid iam rura querar totum subpressa per orbem / Ac iussam seruire famem? Quis castra timenti / 320 Nescit mixta foro, gladii cum triste micantes / Iudicium insolita trepidum cinxere corona, / Atque auso medias perrumpere milite leges, / Pompeiana reum clausurunt signa Milonem? / Nunc quoque ne lassum teneat priuata senectus, / 325 Bella nefanda parat suetus ciuilibus armis / Et docilis Sullam scelerum uicisse magistrum;". A tooltip is visible over the word "micantes", showing the following information: "320 micantes Z<sup>1</sup> P m u", "Paris, Bibliothèque Nationale de France, lat. 7502 (saec. X)", and "Colbertinus sive Parisinus Bibl. Nat. Lat. 7502".

Figura 3. Un esempio di front-end dei metadati dei manoscritti in *Musisque Deoque*.

Ponendo a confronto *front-end* e *back-end*, osserviamo che “Paris” è il valore di *Settlement*, “Bibliothèque Nationale de France” quello del *Repository*, “lat. 7502” è la segnatura (*Shelfmark*) e “(saec. X)” è il valore di *Date*; nella seconda dicitura (quella in colore nero), “Colbertinus sive Parisinus Bibl. Nat. Lat. 7502” è il valore di *Name* seguito dalla segnatura del manoscritto e, come già accennato, corrisponde ai metadati disponibili prima dell'intervento di riorganizzazione. La scelta di non obliterare la precedente metadazione non è casuale: essa consente, sul piano analogico, di rendere note all'utente le (eventuali) varie denominazioni del testimone e, sul piano della risorsa digitale, di lasciare traccia della “storia” del processo di metadazione stesso. Infatti, poiché, stando alle ottime osservazioni di Nicola Barbuti,<sup>4</sup> le entità digitali attualmente esistenti devono essere considerate *entità culturali digitali* (e quindi parte del *Digital Cultural Heritage* o DCH), cioè una «manifestazione ed espressione della storia evolutiva delle comunità del Ventunesimo secolo»,<sup>5</sup> occorre progettare i metadati «tenendo conto non solo delle esigenze di fruizione dei dati nell'immediato, ma soprattutto dei bisogni cognitivi e informativi di chi, un domani, li avrà come fonti principali (se non uniche) alle quali attingere per conoscere gli attuali processi di digitalizzazione e creazione digitale».<sup>6</sup>

Ma torniamo alla visualizzazione delle informazioni sui manoscritti. Se l'utente, invece che passare il cursore, clicca sul *siglum* del ms. in apparato, compare una scheda che riporta i valori di *Settlement*, *Repository*, *Shelfmark* e *Date* (nel caso del ms. Q, “Paris, Bibliothèque Nationale de France, lat. 7900A [saec. X]”) e, sotto, il *permalink* al sito della biblioteca in cui è conservata la fonte e, ove possibile, il *permalink* alla scheda dettagliata della fonte stessa. Al di sotto dei *permalink* compare – ed è questa una novità assoluta – l'elenco di tutte le opere (*Testi*) che, disponendo di apparato in *MQDQ*, sono trasmesse dal testimone selezionato (fig. 4):

<sup>3</sup> L'edizione digitale è a c. di A. Cozzolino (2010); l'inserimento e il controllo dei dati è a c. di P. Marzano; l'edizione critica di riferimento è Bourgery – Ponchont 1967.

<sup>4</sup> Barbuti 2019.

<sup>5</sup> Barbuti 2019, p. 123.

<sup>6</sup> Barbuti 2019, p. 126.



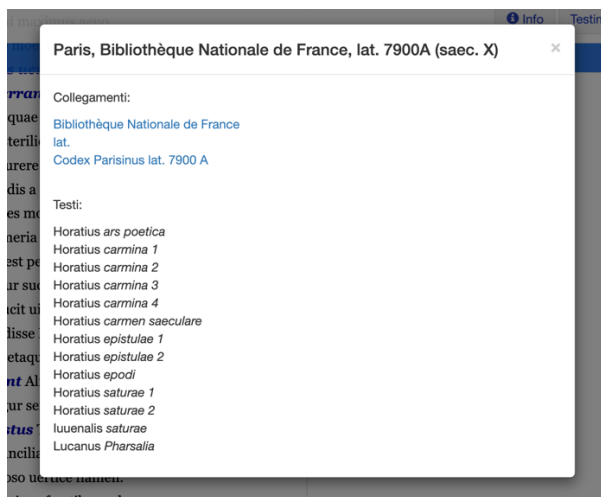


Figura 4. Un esempio di scheda riportante i metadati dei testimoni in *Musisque Deoque*.

L'elenco completo delle fonti antiche è consultabile anche dal menù *Testimoni*. Esso consente all'utente di ottenere in un colpo solo le principali informazioni utili sul fronte bibliografico e critico-testuale. Il menù mostra, infatti, il titolo e la sezione dell'opera (e.g., *Pharsalia*, 1), gli estremi bibliografici dell'edizione critica di riferimento (*Testo base di riferimento*), il nome del curatore dell'edizione digitale (*Cura dell'edizione digitale*) e di chi ha inserito e controllato i dati (*Inserimento e controllo dei dati*). Sul fronte critico-testuale, l'utente visualizza l'elenco completo dei manoscritti che riportano l'opera. Cliccando sui *sigla*, per ogni manoscritto compaiono nuovamente le informazioni essenziali per la conoscenza della fonte e l'elenco delle opere che, essendo dotate di apparato in *MQDQ*, sono trasmesse dal testimone selezionato (come in fig. 3).

#### 4. RISULTATI

I lavori di aggiornamento e di riorganizzazione dei testimoni presenti in *Musisque Deoque* hanno implementato notevolmente l'archivio digitale sotto il profilo quantitativo e qualitativo. In poco più di dieci anni (dal 2012 al 2023) la quantità dei dati è stata considerevolmente accresciuta, incrementando il numero dei testimoni (manoscritti e non) di circa 4700 unità. Inoltre, le possibilità di collegamento tra testi e testimoni sono state moltiplicate.

In ottemperanza al principio FAIR F.2 (*data are described by rich metadata*), l'apparato dei metadati è stato implementato con l'aggiunta dei campi relativi a datazione, dicitura latina e *permalink* alla digitalizzazione del testimone. Poiché in passato i dati e i metadati erano stati inseriti senza seguire precise linee guida, gli interventi più recenti hanno provveduto a eliminare le disomogeneità. In particolare, l'inserimento delle informazioni sulla datazione, prima lasciato al libero arbitrio del singolo editore, è stato disciplinato secondo un protocollo operativo comune e i *link* sono stati implementati in *permalink*, con una metodologia rispettosa delle buone pratiche di uso comune.

Le ridondanze e le ripetizioni sono state evitate. Nel caso delle nomenclature dei testimoni, si è deciso però di mantenere i metadati relativi alla fonte inseriti precedentemente rispetto all'intervento di riorganizzazione: ometterli avrebbe significato obliterare preziose informazioni sul dato e cancellare ogni traccia del processo di metadazione, rischiando di provocare, in prospettiva futura, la perdita di parte della memoria storica relativa all'entità analogica e agli attuali processi di digitalizzazione e creazione digitale.

Tra le novità di maggior rilievo vi è l'aggiunta dell'elenco di tutte le opere che, essendo dotate di apparato in *MQDQ*, sono trasmesse dal singolo testimone. Il fatto che l'elenco si trovi sia nel menù che compare cliccando sul *siglum* di ciascun manoscritto in apparato sia in quello visualizzabile dall'etichetta *Testimoni* crea una piccola ridondanza, la quale però è stata necessaria per evitare un'eccessiva granularità, ossia per collocare il testimone nel suo contesto analogico e digitale. In conclusione, a seguito della ristrutturazione dei metadati, *MQDQ* si presenta come uno strumento informatico e filologico che garantisce accessibilità, costante consultazione (secondo i principi [FAIR](#)) e scientificità nella creazione di collegamenti alle collezioni digitali delle più importanti biblioteche europee e nordamericane. La metadazione consente non solo il reperimento delle informazioni sul dato, ma anche di quelle sulla "storia" del processo di metadazione stesso.

#### BIBLIOGRAFIA

[1] Barbuti, Nicola. «Ripensare i formati, ripensare i metadati: prove "tecniche" di conservazione digitale». *Umanistica digitale* 5 (2019): 121-138. DOI: <https://doi.org/10.6092/issn.2532-8816/9055>

[2] Boschetti, Federico; Del Grosso, Angelo Mario; Spinazzè, Linda. «La galassia Musisque Deoque: storia e prospettive». In *Paulo maiora canamus*. Raccolta di studi per Paolo Mastandrea, a cura di M. Venuti e M. Manca, 405-419. Venezia: Edizioni Ca' Foscari, 2021.

[3] Bourguery, Abel - Ponchont, Max. *Lucain: La guerre civile: La Pharsale*, tome 1, livres I-V. Paris: Les Belles Lettres, 1967.

[4] Mastandrea, Paolo. «Archivi elettronici di poesia latina e opzioni multiple di ricerca testuale». *Semicerchio: rivista di poesia comparata* 53.2 (2015): 60-69.

[5] Spinazzè, Linda. «Risalire alle fonti: dall'edizione Musisque Deoque ai testimoni manoscritti». In *Nuovi archivi e mezzi d'analisi per i testi poetici: lavori del progetto Musisque Deoque*. Venezia 21-23 giugno 2010, a cura di P. Mastandrea e L. Spinazzè, 59-71. Amsterdam: Adolf M. Hakkert editore, 2011.

# Storytelling with Linked Open Data

Giulia Renda<sup>1</sup>, Marilena Daquino<sup>2</sup>

<sup>1</sup> University of Bologna, Italy – giulia.renda3@unibo.it

<sup>2</sup> University of Bologna, Italy – marilena.daquino2@unibo.it

## ABSTRACT

Data visualisation techniques in combination with storytelling strategies can help to make more explicit the complex network of information and relationships that characterises Linked Open Data (LOD). However, existing LOD visualisation tools show that the narrative element in the graphical representation is overlooked or missing altogether. In this poster we present MELODY, a web application to query, visualise, present and disseminate LOD as data stories.

## KEYWORDS

Linked Open Data; data visualisation; storytelling; ux design

## 1. INTRODUCTION

In recent years, the potential offered by Semantic Web technologies is being leveraged by researchers and cultural institutions, which often release their cultural heritage data in the form of Linked Open Data (LOD). Yet users have little opportunity to explore such data but through the query interfaces offered by the providers. The result is a barrier to cultural dissemination. In this context, stories can play an important role. Scholars [3] claim that stories can serve as an interpretive framework to convey the values and meaning of a culture, but also to express very complex information to the user in a more compact way [6][7]. For the success of both open data business models and quantitative analysis methods in the digital humanities, it is desirable to have user-friendly and customisable interfaces that allow non-expert users to interact with data, create stories and communicate findings effectively. In this poster, we introduce MELODY<sup>1</sup>, an open source tool and online service for creating data stories based on LOD. Currently, MELODY offers three templates: a dashboard, an interactive text search, and a map. We introduce the methodology we implemented to gather requirements, including an initial mapping between interaction patterns and competency questions derived from a case study in the music heritage domain, i.e. Polifonia, and we present the main features of the platform.

## 2. RELATED WORKS

Authors in [4] thoroughly analysed 77 Linked Data visualisation tools and demonstrated that there is a lack of adequate solutions for LOD-based storytelling that provide full support to users who want to explore Linked Data and gain knowledge from it. Four main issues can be identified. As highlighted in [1], (1) most traditional exploration and visualisation systems work offline and are limited to accessing static sets of pre-processed small datasets (RDFShape [5]). Furthermore, (2) access to these datasets may be limited to predefined datasets that are not accessible via SPARQL endpoints or API requests (Graphless [11]). Even if tools can explore many datasets, (3) they come with simple statistics (H-BOLD [8]) or present RDF data only in tabular form (RDFSurveyor [12]). Tools that are more concerned with usability or that do not face the problems mentioned above still (4) do not offer the user the possibility to present data in an author-controlled narrative (SynopsViz [2]). With MELODY we want to tackle these problems by offering a reusable interface for storytelling that is entirely based on Linked Open Data.

## 3. FROM ONTOLOGY DESIGN TO STORYTELLING INTERFACES

In Polifonia<sup>2</sup>, the development of a Linked Open Dataset requires knowledge engineers and domain experts to agree on a set of Competency Questions (CQs) that are building blocks for the ontology design process [9]. In designing MELODY, we relied on ~220 CQs<sup>3</sup> collected from historians, musicologists and cultural heritage professionals (who are also early adopters of MELODY). The CQs are organised into 19 personas and 28 stories [10], each representing typical users and associated behaviours. We annotated the CQs with entities, relations, and potential UI/UX search patterns<sup>4</sup>. From aggregate

---

<sup>1</sup> <https://projects.dharc.unibo.it/melody>

<sup>2</sup> <https://polifonia-project.eu/>

<sup>3</sup> <https://github.com/polifonia-project/stories>

<sup>4</sup> [https://github.com/polifonia-project/web\\_portal/tree/main/analysis](https://github.com/polifonia-project/web_portal/tree/main/analysis)

data analysis, we estimated data types (bibliographic, linguistic, music), data patterns (e.g. distribution of CQs in terms of people/places/works/events, etc.) and interaction patterns (e.g. CQs returning result lists/charts/maps). Then we analysed the stories individually and identified driving CQs, i.e. the ones that move all subsequent research, and we assigned a main pattern UI /UX to each story. The results are four interaction patterns (three of which are available as templates in MELODY), namely: a dashboard-alike interface, an interactive text search, a geographic map and an evolutionary graph.

#### 4. CREATE A DATA STORY WITH MELODY

Make mE a Linked Open Data storY, MELODY<sup>5</sup> is a web application for querying any SPARQL endpoint and sharing data stories, i.e. web-enabled documents with data visualisations and curated narratives that effectively communicate content and ideas.

To create a story, users are guided in a walkthrough (Figures 1 and 2): (1) specify the URL of a SPARQL endpoint, (2) choose a graphic template, and (3) query data sources through SPARQL queries to populate the story with different graphs, contents, and interactive searches. There are currently three interaction patterns: a dashboard-like interface where curated text can be associated to charts to create a blog post; an interactive text search where results of a search can be linked to new searches to create a guided tour of the dataset; a geographic map where data points can be filtered out, and contextual statistical information can be visualised.

Users are guided through a WYSIWYG interface where they can start creating the story by selecting among different components provided. Figure 3 shows how this type of interface can instantly query and preview the results: as soon as a functioning SPARQL query is provided, its results are displayed according to the chosen type of visualisation. The resulting data stories can be exported (PDF/HTML) and published, either on the MELODY website or in a catalogue<sup>6</sup>, and easily shared online.

MELODY provides an authentication system via GitHub that allows users to create and interact with content. We can distinguish three types of users, namely: (1) a user belonging to the organisation that sets up the platform (a Polifonia collaborator in the use case example), who can create and publish content directly on MELODY and can edit data stories at any time; (2) an external user, who can only publish on a separate website (the catalogue mentioned above); (3) an unauthenticated user who cannot store the data story anywhere, but only download it. Figure 4 shows an example of how the data story or the single component (a barchart in this case) can be exported.

Although MELODY was developed to support pilot projects that are part of Polifonia, it can be reused by any user who wants to explore any data available via a SPARQL endpoint. The concept of reuse is certainly related to that of extensibility. Indeed, the architecture and extensible templates facilitate the addition of new functionalities.

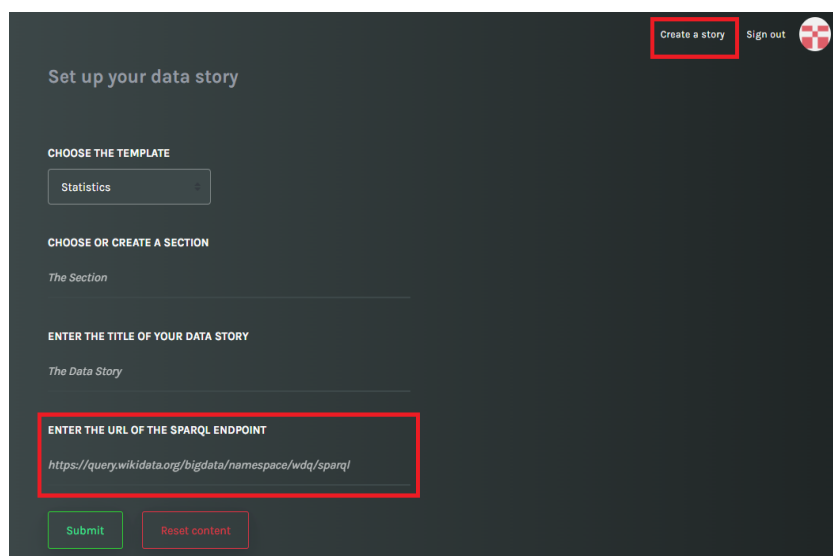


Figure 1. Setup form to start creating a data story.

<sup>5</sup> <https://polifonia-project.github.io/dashboard/>

<sup>6</sup> <https://projects.dharc.unibo.it/melody>; <https://melody-data.github.io/stories/>

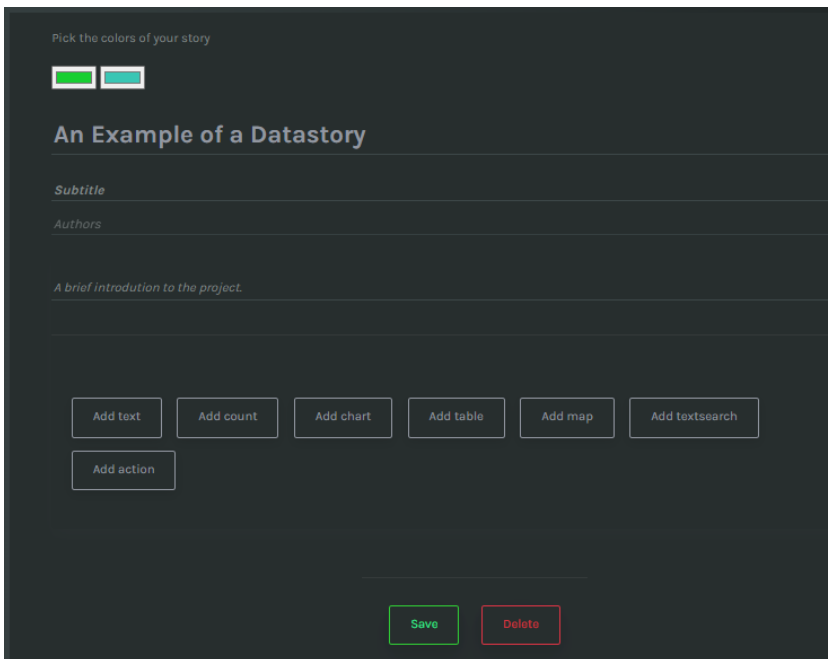


Figure 2. The initial WYSIWYG interface with the list of components.

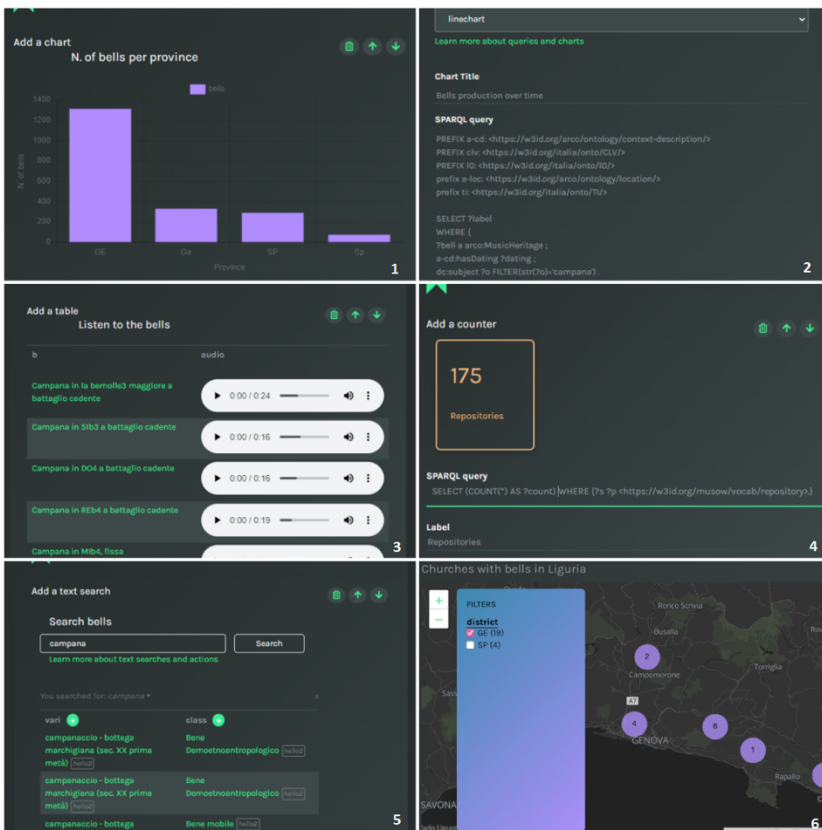


Figure 3. An overview of the components: a chart preview (1), the form for the chart (2), a table preview with audio files (3), a count preview (4), an example of textsearch (5), the preview of the map with filters (6).

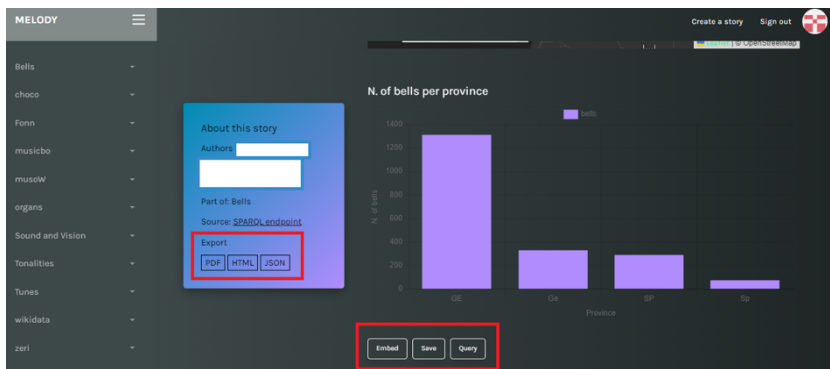


Figure 4. Buttons to export data stories, configuration files, and single components.

## 5. ACKNOWLEDGMENTS

This work has received funding from the European Union’s Horizon 2020 research and innovation programme (Polifonia, G.A. 101004746).

## REFERENCES

- [1] Bikakis, Nikos, and Timos Sellis. ‘Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art’. ArXiv:1601.08059 [Cs], 29 January 2016. <http://arxiv.org/abs/1601.08059>.
- [2] Bikakis, Nikos, Melina Skourla, and George Papastefanatos. ‘Rdf:SynopsViz - A Framework for Hierarchical Linked Data Visual Exploration and Analysis’. arXiv, 27 June 2017. <https://doi.org/10.48550/arXiv.1408.3148>.
- [3] Bruner, Jerome. ‘The Narrative Construction of Reality’. *Critical Inquiry* 18, no. 1 (1991): 1–21.
- [4] 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>.
- [5] Gayo, Jose Emilio Labra, Daniel Fernández-Álvarez, and Herminio García-González. ‘RDFShape: An RDF Playground Based on Shapes’. In *SEMWEB*, 2018.
- [6] Gershon, Nahum, and Ward Page. ‘What Storytelling Can Do for Information Visualization’. *Communications of the ACM* 44, no. 8 (1 August 2001): 31–37. <https://doi.org/10.1145/381641.381653>.
- [7] Lombardo, Vincenzo, and Rossana Damiano. ‘Storytelling on Mobile Devices for Cultural Heritage’. *New Review of Hypermedia and Multimedia* 18, no. 1–2 (1 March 2012): 11–35. <https://doi.org/10.1080/13614568.2012.617846>.
- [8] Po, L., and D. Malvezzi. ‘High-Level Visualization over Big Linked Data’, Vol. 2180, 2018.
- [9] Presutti, Valentina, Enrico Daga, Aldo Gangemi and Eva Blomqvist. “eXtreme Design with Content Ontology Design Patterns.” *WOP* (2009).
- [10] Pruitt, John, and Tamara Adlin. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Elsevier, 2006.
- [11] Santana-Pérez, Idafen. ‘Graphless: Using Statistical Analysis and Heuristics for Visualizing Large Datasets’. In *VOILA@ISWC*, 2018.
- [12] Vega-Gorgojo, Guillermo, Laura Slaughter, Bjørn Marius Von Zernichow, Nikolay Nikolov, and Dumitru Roman. ‘Linked Data Exploration With RDF Surveyor’. *IEEE Access* 7 (2019): 172199–213. <https://doi.org/10.1109/ACCESS.2019.2956345>.

# From ontology design to user-centred interfaces for music heritage

Giulia Renda<sup>1</sup>, Marco Grasso<sup>2</sup>, Marilena Daquino<sup>3</sup>

<sup>1</sup> University of Bologna, Italy – giulia.renda3@unibo.it

<sup>2</sup> University of Bologna, Italy – marco.grasso7@unibo.it

<sup>3</sup> University of Bologna, Italy – marilena.daquino2@unibo.it

## ABSTRACT

In this article we investigate the bridge between ontology design and UI/UX design methodologies to assist designers in prototyping web applications for information seeking purposes. We briefly review the state of the art in ontology design and UI/UX methodologies, then we illustrate our approach applied to a case study in the music heritage domain.

## KEYWORDS

Music heritage; ontology design; ux design; generous interfaces

## 1. INTRODUCTION

Polifonia<sup>1</sup> is a European project that aims at connecting resources in the music heritage and to engage with experts and the general public. Ten pilot projects have been designed for the purpose, wherein scholars collect data sources (e.g. texts, audio files), extract information, and transform data into Linked Open Data (LOD) to populate a knowledge graph. The knowledge extraction is driven by competency questions, which also guide the ontology design. The ultimate goal of the knowledge graph is to be leveraged in one or more web applications.

However, eliciting data/user requirements, ensuring completeness, and selecting the right approach – e.g. modular, bottom-up, user-centred – to develop User Interfaces (UI) and Experience (UX) based on domain requirements is challenging. While several efforts have been made to integrate methods of Human-Computer Interaction (HCI) into ontology design methodologies, to the best of our knowledge, there is no overall ontology-driven methodology for developing multi-purpose web solutions. In particular, how can we create interfaces for presenting data characterised by a broad and diverse scope leveraging Linked Open Data and domain ontologies? How do we ensure interfaces are specialised enough to answer complex questions but are usable by stakeholders with different backgrounds?

In this article we investigate the bridge between ontology design and UI/UX design methodologies to assist designers in prototyping web applications for information seeking purposes. We briefly review the state of the art in ontology design and UI/UX methodologies, then we illustrate our approach applied to the Polifonia case study. The methodology can be reused in similar contexts where cultural heritage data are disseminated on the web.

## 2. STATE OF THE ART

In the Semantic Web, the structure of SPARQL queries is closely related to ontologies. This means that users' information needs can be mapped to both ontology requirements and interfaces requirements. Therefore, we would expect that ontology design methods are closely related to UI/UX methodologies. Several ontology design methodologies have indeed adopted tools from HCI. Such methods favour a bottom-up approach to elicit requirements and rely on the intervention of domain experts in (1) defining the domain space and vocabulary, (2) outline motivating scenarios, and (3) extracting requirements in the form of natural language Competency Questions (CQs) [8] [14] [21] [5].

The eXtreme Design (XD) methodology [16] [3] prescribes practices for capturing goals, interests, and tasks from stakeholders, grouping them under umbrella categories, i.e. personas [10]. Unlike other methodologies, XD also captures research journeys, expectations, and priority levels. Since personas also drive the definition of users' behaviours search interfaces, the knowledge acquisition process is closer to current practices in UI/UX design. For these reasons, XD is a good candidate for a seamless integration into UI/UX design processes.

When mapped to UI elements, interactive behaviours can lead to two types of search paths. On the one hand, exploratory journeys can focus on retaining the user, breaking down the information in small chunks to reduce the cognitive load [11]. On the other hand, preventing the user from seeing the “whole picture” could be disorienting. Studies in Information

---

<sup>1</sup> <https://polifonia-project.eu/>



Science suggest that “third generation” information systems should first filter out data of interest and then apply data analysis and knowledge discovery tools on the target [9]. To meet this call, scholars advocate for more generous interfaces [22], leading to an approach based on “overview first, zoom and filter, then details on demand” [19].

To reconcile these two different perspectives, the user interface must be able to support multiple tasks and user journeys [1]. In this respect, stories are powerful tools for designing experiences, as they present facts connected by causal relationships, and help to formulate users’ motivational aspects or to describe unforeseen situations [7].

Design Thinking (DT) [18] [2] [6] [13] is a user-centred approach to problem solving, based on a hypothesis-driven, abductive and dialectical approach to map requirements to design ideas. Previous studies have shown that DT effectively improves the quality of the ideas generated and reduces the risk of failure [12]. It consists of six phases: empathise, define, ideate, prototype, test, and implement. In the data collection phase, various methods are used, including personas, stories, stakeholder and user journey maps [4]. Several studies [17] [20] [15] have attempted to incorporate TD into specific aspects of ontologies creation. Results mostly provide formal definitions of HCI and DT methods, but do not inform us on how to leverage real-world domain ontologies in the DT process. In this work, we aim at filling this gap, suggesting the application of methods and analyses widely recognised as tools of the DT methodology directly to the domain ontologies during the ontology design phase

### 3. METHODOLOGY

We introduce a modular workflow harmonising eXtreme Design and Design Thinking, that reuses content/user requirements in UI/UX design. We identify nine stages, grouped in three main activities (Figure 1), namely:

- **Ontology design.** (1) The ontology design team outlines personas and groups them, (2) writes one or more stories for each persona, and (3) extracts competency questions.
- **User interfaces design.** (4) The web development team (us) defines the most important CQs, which serve as drivers for all others, and (5) group remaining questions into meaningful clusters. (6) We analyse drivers and clusters and select appropriate visualisation types for the reference data.
- **User experience design.** (7) The web development team outlines interaction patterns (via competitive analysis or focus groups), (8) selects appropriate solutions to deploy, and (9) performs user testing validation.

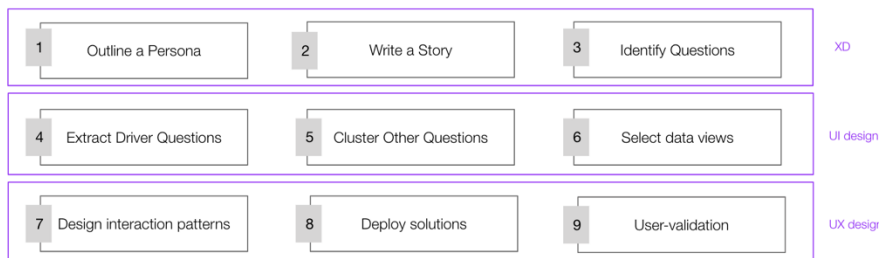


Figure 1. Overview of the methodology.

### 4. CASE STUDY: DESIGNING POLIFONIA INTERFACES

The Polifonia ontology network describes music sources, performances, instruments, and music features. The Polifonia ecosystem currently includes 9 datasets, for which 19 personas, 28 stories, and 240 competency questions have been identified<sup>2</sup>. Following our workflow, we have identified four web applications to be developed, namely: musoW<sup>3</sup>, a filter-based catalogue of music data on the web, targeted to music professionals; MELODY<sup>4</sup>, a web editor of data visualisations and stories, targeted to music domain experts; Corpus<sup>5</sup>, to perform linguistic analysis over a vast corpus of music-related text sources; and the Polifonia Web portal (in progress), to present data to lay users according to several strategies. An example of the workflow applied to Polifonia for generating musoW is the following.

1. **Persona building.** Laurent is a music journalist. He searches the web for new music sources, but he does not have sufficient technical skills to perform these searches systematically.

<sup>2</sup> <https://github.com/polifonia-project/stories>

<sup>3</sup> <https://projects.dharc.unibo.it/musow/>

<sup>4</sup> <https://projects.dharc.unibo.it/melody/>

<sup>5</sup> <https://polifonia.disi.unibo.it/corpus/>

2. **User Story.** Laurent publishes a weekly newsletter in which he summarises his findings in the music industry. To gather information, he created a text document with a list of music resources that he checks regularly. Unfortunately, limited searches can be done on the document. Therefore, he would like to have access to an online catalogue that allows more sophisticated filtering options.
3. **Competency Questions.** The following questions were extracted from the story:
  - a. CQ1: Can I search for musical content by applying filters (genre, period ...)?
  - b. CQ2: What types of resources can I find?
  - c. CQ3: Is the music resource X complete or incomplete?
  - d. CQ4: Is a dataset attached to resource X?
  - e. CQ5: Can I add resources as a user?
  - f. CQ6: How can I share what I find on the site?
 Some preliminary requirements can already be detected from the questions, namely:
  - Data requirements: genre and time (CQ1), type (CQ2) and availability (CQ4).
  - Functional requirements: filters (CQ1), completeness (CQ3), crowdsourcing (CQ5), share (CQ6).
4. **Driver questions.** We identify CQ1 as the driving question, as it defines the problem space (*can I search for*), identifies the reference entity (*musical content*), and suggests how to visualise the data (*filters*).
5. **Clusters of questions.** Other CQs can be split in two groups: CQ2-4 address context information of the main entity (*type, complete/incomplete, dataset*); CQ5-6 address actions to be performed on the data (*add, share*).
6. **Select Data views.** The driver question matches a specific type of data visualisation, i.e. a filter-based exploration of resources ordered by relevance, hence no further analysis is needed.
7. **Interaction patterns.** In order to develop generous interfaces, we review web applications that present similar tasks, and then we map UI patterns to CQs:
  - a. CQ1: group resources under categories and show the counting for each category, to give an overview.
  - b. CQ2-CQ4: show lists of resources for each category on demand.
  - c. CQ5, CQ6: provide specialised operations when browsing the record of a resource.
8. **Deploy solutions.** We check whether there are other personas with similar information requirements (step 1). Since no other personas have similar requests and the call for action is rather specific, we continue with the development of a bespoke solution, i.e. musoW.
9. **User test.** musoW was validated in focus groups with stakeholders, competitors, and project partners.

Laurent is the only persona that required a dedicated, specialised application for browsing music resources on the web. Other personas are either scholars with very specific research questions (for which we developed MELODY and the Corpus) or lay users, who do not have a specific task guiding their exploration (for which we develop a web portal). Iteratively applying the workflow to each persona may be time-consuming and does not ensure results are representative of the whole picture (rather, the result is simply going to be the sum of all requirements). Therefore, when analysing the remaining 18 personas, step 6 (Select Data views) is extended with a distant reading approach, performing an exploratory analysis of CQs. We manually annotated CQs in an online table<sup>6</sup> (see an example in Figure 2) with scope, classes, ontology patterns, and expected type of result (e.g. list, map, single result, explanation).

CQ ID	exp.	CQ	Bibl. data	Music data	Lingui:	User da	Main entities	Additional entities/props
Carolina1-CQ1	-	Where was a musical composition performed?	yes	no	no	no	Musical performance: musical composition	
Carolina1-CQ2	-	In which buildings was a musical composition performed?	yes	no	no	no	Musical perform; Musical performance: bui	
Carolina1-CQ3	-	Where was a musical composition performed for the first time?	yes	no	no	no	Musical perform; Musical performance: dat	

Figure 2. Manual annotation of Competency Questions.

We analysed results to grasp an overview of priorities, data patterns, and user journeys. The preliminary analysis of CQs is available online as a Jupyter notebook<sup>7</sup>. In detail, we identified three categories of data and estimated their coverage:

- bibliographic data (on music works, historical events, composers) covered by 70% of CQs;
- structured music data (melody, harmony, rhythm), covered by 34% of CQs;

<sup>6</sup> <https://docs.google.com/spreadsheets/d/16hr2fFTc4VUQHob0ALTu1TJ95xtyGvmWvly2yxOZFcM>

<sup>7</sup> [https://colab.research.google.com/drive/17I\\_3yjo2XoDDw6OTLvtAOE1PjOZ5xgmg?usp=sharing](https://colab.research.google.com/drive/17I_3yjo2XoDDw6OTLvtAOE1PjOZ5xgmg?usp=sharing)

- linguistic, full-text data (emotions, song lyrics), covered by 30% of CQs. Among these, 77% rely also on bibliographic data, and 19% also on music data. Only 5% of CQs require all three types of data.

We assume we can identify priorities as the most representative data requirements (bibliographic) and estimate the complexity of services to be implemented as those that satisfy niche areas (musicologists and linguists). Secondly, we analysed entities and ontology patterns. To this end, we identified an input, intermediate, and output entities for each CQ and we used a Sankey diagram to visualise journeys (Figure 3). For instance Carolina-CQ1 (Figure 2) “Where was a musical work performed” has “Music Work” as input, “Musical performance” as intermediate and “Place” as output.

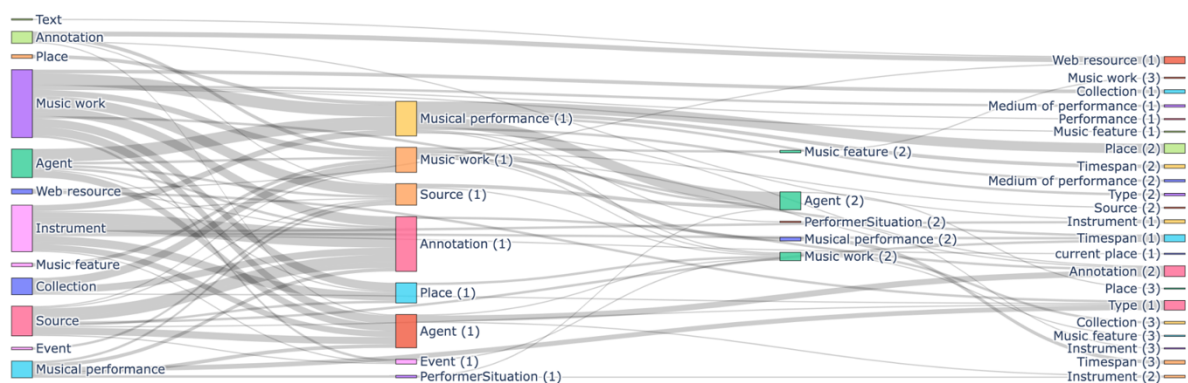


Figure 3. Data patterns and user journeys.

Music works, agents, and sources are the main access points to the knowledge graph, intermediate aspects address more or less technical aspects, such as annotations (music and linguistic data), agents, and events, and final outputs are rather diverse. Again, we assume that most recurring patterns are priorities for developing UI components and journeys.

Step 7 (Interaction patterns) is performed via a competitive analysis, i.e. a user study wherein participants were requested to validate two similar web applications leveraging different UX strategies more/less similar to generous interfaces<sup>8</sup>. Results of the survey drove the final definition of UI components and their composition in the web page.

Finally, step 9 (User validation) is performed via another user study, this time devoted to co-design aspects<sup>9</sup>. In particular, users are asked to answer questions on how they would like or expect a website for music data recommendation to look like and behave. Users are asked to imagine themselves in a scenario they are comfortable with (e.g. “you are at home and you want to discover new music”) and to describe their research process and expectations. It’s worth noting that they do not see a website to evaluate. Results of the survey are matched against decisions already taken in step 7, which provide us, in a reverse-engineering fashion, with an evaluation of the expected user satisfaction. We believe it is fair to assume that the validation of the results produced by using our methodology can be inherited by the methodology itself.

## 5. CONCLUSION

We defined a workflow that leverages ontology requirements in UI/UX design to (1) develop ideas and prototypes that match data requirements, (2) have the UI/UX design iteratively informing and revising ontology requirements. The usage of exploratory analysis on Competency Questions gave us an overview of data requirements and a clear definition of priorities. Data patterns allow us to estimate types of content interaction and their relevance. Two user studies (focused respectively on competitive analysis and co-design techniques) help us to calibrate services and expectations of a wide range of users, including experts and lay people. Our preliminary results lead us to justify two strong assumptions, namely: (1) similar CQs can be grouped by type of interaction pattern; (2) entities that are relevant to a large number of CQs are also likely to be relevant to a wide range of users. As a consequence, we can apply our workflow to a much smaller number of groupings of CQs (instead of over each persona), therefore preventing time-consuming activities.

In conclusion, binding ontology design requirements to UI/UX design choices revealed being a good solution to tackle common issues in projects dedicated to the dissemination of cultural heritage data. Results on the case study were successful, and preliminarily validated the goodness of our approach, which was applied when designing four applications having different goals (a catalogue, an authoring platform, a linguistic corpus interface, and a web portal for engaging lay

<sup>8</sup> <https://docs.google.com/spreadsheets/d/1Q1Byjk9oAutD9yDk2Gaig-5VYqrSbRO2SCu8Lz9MJf0>

<sup>9</sup> [https://docs.google.com/spreadsheets/d/1Ky83VoAUJRcMYSLlgaedw5QYH\\_s5KCNdPicUyWFq320](https://docs.google.com/spreadsheets/d/1Ky83VoAUJRcMYSLlgaedw5QYH_s5KCNdPicUyWFq320)

users). In future works we plan to test our methodology in projects with a different scope, in order to validate the reusability of our methodology in contexts different from the one in which it has been developed.

## 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] Bevan, Nigel, Jim Carter, Jonathan Earthy, Thomas Geis, and Susan Harker. 'New ISO Standards for Usability, Usability Reports and Usability Measures'. In *Human-Computer Interaction. Theory, Design, Development and Practice: 18th International Conference, HCI International 2016, Toronto, ON, Canada, July 17-22, 2016. Proceedings, Part I* 18, 268–78. Springer, 2016.
- [2] Bevan, Nigel, Jim Carter, Jonathan Earthy, Thomas Geis, and Susan Harker. 'New ISO Standards for Usability, Usability Reports and Usability Measures'. In *Human-Computer Interaction. Theory, Design, Development and Practice: 18th International Conference, HCI International 2016, Toronto, ON, Canada, July 17-22, 2016. Proceedings, Part I* 18, 268–78. Springer, 2016.
- [3] Brown, Tim and others. 'Design Thinking'. *Harvard Business Review* 86, no. 6 (2008): 84.
- [4] 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.
- [5] Chasanidou, Dimitra, Andrea Alessandro Gasparini, and Eunji Lee. 'Design Thinking Methods and Tools for Innovation'. In *Design, User Experience, and Usability: Design Discourse: 4th International Conference, DUXU 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings, Part I*, 12–23. Springer, 2015.
- [6] Cota, Giuseppe, Marilena Daquino, and Gian Luca Pozzato. *Applications and Practices in Ontology Design, Extraction, and Reasoning*. Vol. 49. IOS Press, 2020.
- [7] Dorst, Kees. 'The Core of 'Design Thinking' and Its Application'. *Design Studies* 32, no. 6 (2011): 521–32.
- [8] Gruen, Dan, Thyra Rauch, Sarah Redpath, and Stefan Ruettinger. 'The Use of Stories in User Experience Design'. *International Journal of Human-Computer Interaction* 14, no. 3–4 (2002): 503–34.
- [9] Gruninger, Michael. 'Methodology for the Design and Evaluation of Ontologies'. In *Proc. IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.
- [10] Hyvönen, Eero. 'Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery'. *Semantic Web* 11, no. 1 (2020): 187–93.
- [11] Junior, Plinio Thomaz Aquino, and Lucia Vilela Leite Filgueiras. 'User Modeling with Personas'. In *Proceedings of the 2005 Latin American Conference on Human-Computer Interaction*, 277–82, 2005.
- [12] Kiruthika, Jay, Souheil Khaddaj, Darrel Greenhill, and Jarek Francik. 'User Experience Design in Web Applications'. In *2016 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)*, 642–46. IEEE, 2016.
- [13] Liedtka, Jeanne. 'Evaluating the Impact of Design Thinking in Action'. In *Academy of Management Proceedings*, 2017:10264. Academy of Management Briarcliff Manor, NY 10510, 2017.
- [14] Norman, Don. *The Design of Everyday Things: Revised and Expanded Edition*. Basic books, 2013.
- [15] Noy, Natalya Fridman, and Carole D Hafner. 'The State of the Art in Ontology Design: A Survey and Comparative Review'. *AI Magazine* 18, no. 3 (1997): 53–53.
- [16] Pileggi, Salvatore F. 'Knowledge Interoperability and Re-Use in Empathy Mapping: An Ontological Approach'. *Expert Systems with Applications* 180 (2021): 115065.
- [17] Presutti, Valentina, Enrico Daga, Aldo Gangemi, and Eva Blomqvist. 'EXtreme Design with Content Ontology Design Patterns'. In *Proc. Workshop on Ontology Patterns*, 83–97, 2009.
- [18] Ramaprasad, Arkalud, and Thant Syn. 'Design Thinking and Evaluation Using an Ontology'. In *Design Science: Perspectives from Europe: European Design Science Symposium, EDSS 2013, Dublin, Ireland, November 21-22, 2013. Revised Selected Papers*, 63–74. Springer, 2014.
- [19] Rowe, Peter G. *Design Thinking*. MIT press, 1991.
- [20] 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. IEEE, 1996.
- [21] Silva, Thiago Rocha, Jean-Luc Hak, and Marco Winckler. 'A Formal Ontology for Describing Interactive Behaviors and Supporting Automated Testing on User Interfaces'. *International Journal of Semantic Computing* 11, no. 04 (2017): 513–39.
- [22] Suárez-Figueroa, Mari Carmen, Asunción Gómez-Pérez, and Mariano Fernandez-Lopez. 'The NeOn Methodology Framework: A Scenario-Based Methodology for Ontology Development'. *Applied Ontology* 10, no. 2 (2015): 107–45.
- [23] Whitelaw, Mitchell. 'Generous Interfaces for Digital Cultural Collections'. *Digital Humanities Quarterly* 009, no. 1 (21 May 2015).

# Un modello open data per l'interazione utente: la digital library Open Memory Apulia

Stefania Riso<sup>1</sup>

<sup>1</sup> Università degli Studi di Bari Aldo Moro, Italia – stefania.riso@uniba.it

## ABSTRACT

Il paper descrive i primi risultati prodotti dalla ricerca e sviluppo di una digital library innovativa, basata su un modello in open data flessibile. Il progetto di ricerca coinvolge collaborativamente i ricercatori del Dipartimento di Ricerca e Innovazione Umanistica DIRIUM dell'Università degli Studi di Bari Aldo Moro e della spin off di UNIBA D.A.BI.MUS. Srl, attiva nella ricerca e sviluppo nei settori dell'innovazione digitale per il patrimonio culturale.

## PAROLE CHIAVE

Digital Library (DL); cultura; innovazione; patrimonio culturale digitale

## 1. INTRODUZIONE

Nel corso degli ultimi trent'anni si assiste a una crescente proliferazione di Digital Libraries dalle caratteristiche differenti che rispecchiano la loro evoluzione concettuale [1], passando da semplici contenitori di riproduzioni virtuali delle collezioni analogiche possedute dalle biblioteche pubbliche e private a erogatori di servizi dei beni librari, fino a divenire luoghi virtuali dove musei, biblioteche e archivi coesistono e interagiscono e le comunità di utenti apprendono, interagiscono e vivono un'esperienza educativa e culturale [2]. Inoltre, già dal 2020 l'UE [3] ha chiesto di rendere i dati narrazioni sulle identità e sull'agire individuale e collettivo delle comunità europee, trasformandoli nelle nuove fonti storiche del presente per le generazioni future, e in questo scenario, il paradigma *Open Data* ha un ruolo determinante nella prospettiva di definire, creare e gestire il nuovo Patrimonio Culturale Digitale [4].

## 2. METODOLOGIA DELLA RICERCA

Un progetto pilota che mira a rispondere alle recenti istanze di innovazione sostenibile nel settore culturale è Open Memory Apulia<sup>1</sup> (Figura 1), una Digital Library in Open Data progettata dal Dipartimento di Ricerca e Innovazione per le Scienze Umanistiche (DIRIUM)<sup>2</sup> e dallo spin-off D.A.BI.MUS. Srl<sup>3</sup>, entrambi dell'Università degli Studi di Bari Aldo Moro<sup>4</sup>, al fine di raccontare storie di interesse e consentire l'accesso, il riutilizzo e la redistribuzione delle risorse digitali storiche, promuovendo nuove forme di partecipazione sociale e l'interazione attiva delle comunità antropiche con un nuovo Patrimonio digitale.

L'architettura della DL è stata sviluppata seguendo principi fondamentali quali [5]: 1. Architettura aperta (deve essere estendibile e interoperabile); 2. Robustezza (deve essere affidabile); 3. Scalabilità (deve essere adattabile); 4. Conservazione (deve garantire un accesso permanente alla raccolta del DL); 5. Modularità (deve garantire una rimodulazione in base alle diverse esigenze e alle nuove tecnologie emergenti); 6. Privacy (deve essere attenta ai diritti di accesso e alle licenze d'uso); 7. Fattibilità (deve essere realizzabile considerando fattori gestionali ed economici); 8. Interattività (deve essere incentrata sull'interazione con l'utente).

Partendo da questi criteri e dalla volontà di superare la visione *service-oriented*, realizzando un'architettura *user-oriented*, la DL è stata progettata utilizzando il DMS CKAN<sup>5</sup> per creare, gestire, valorizzare, conservare e disseminare i dati del patrimonio digitale della regione Puglia, secondo il paradigma Open Data per migliorare l'interazione *user-centered*. La struttura ideata e progettata per la DL si mostra in duplice aspetto: da un lato è user-friendly, consentendo all'utente di raggiungere le risorse digitali attraverso diverse modalità di accesso o indirizzandolo, mediante tag e filtri, nel percorso di ricerca; dall'altro lato, invece, è una realtà complessa in grado di gestire entità complesse (collezioni, singoli oggetti

<sup>1</sup> Open Memory Apulia, Home Page, <http://openmemoryapulia.it> (ultima consultazione 04.05.2023).

<sup>2</sup> Dipartimento di Ricerca e Innovazione Umanistica (DIRIUM), Home Page, <https://www.uniba.it/it/ricerca/dipartimenti/dirium> (ultima consultazione 04.05.2023).

<sup>3</sup> D.A.BI.MUS. S.r.l., Home Page, <https://www.dabimus.com> (ultima consultazione 04.05.2023).

<sup>4</sup> University of Bari Aldo Moro, Home Page, <https://www.uniba.it/it> (ultima consultazione 04.05.2023).

<sup>5</sup> CKAN, The world's leading open source data management system, <https://ckan.org> (ultima consultazione 04.05.2023).



digitali) e diversificate per natura e tipologia, collegate tra loro mediante legami di significato, nonché le relazioni e interazioni che esse generano con altre entità fisico-digitali (utente, operatori, software).

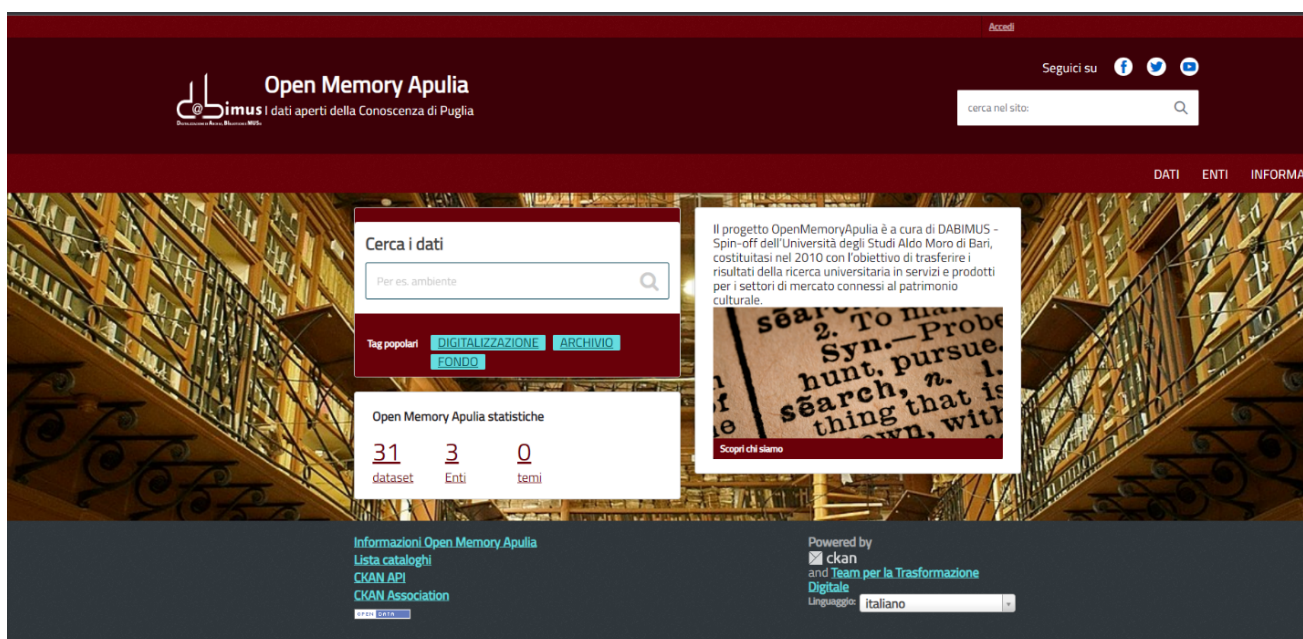


Figura 14. Homepage della Digital Library.

Il progetto Open Memory, pertanto, si è focalizzato su due elementi peculiari che permettono al sistema di gestione della Digital Library di offrire all'utente, interno o esterno al sistema, servizi base per la sottomissione, descrizione, indicizzazione, ricerca, browsing, recupero, preservazione e visualizzazione di oggetti digitali ma, anche, servizi specialistici, come per esempio il controllo degli accessi:

La gestione degli oggetti digitali e dei metadati. Considerando i metadati [6] [7] di una risorsa culturale digitale come particelle atomiche che popolano un ecosistema digitale e trasmettono le informazioni culturali in maniera diacronica nello spazio e nel tempo, è stato progettato un tracciato di metadati strutturato in tag associati ai metadati per renderli semanticamente comprensibili, con l'obiettivo di registrare informazioni relative sia agli artefatti culturali analogici sia a oggetti digitali, inclusi i metadati che descrivono la provenienza e il ciclo di vita dei dati, tra cui: contenuto, utenti, strumenti, qualità, regole di uso e riuso e architettura. Inoltre, il tracciato dei metadati si presenta flessibile permettendo di compilare o disattivare i suddetti tag, adattando il numero di metadati e le descrizioni alle diverse tipologie di patrimonio analogico digitalizzato, aggiungendo ulteriori informazioni o variandone l'estensione e aggiornando in tempo reale i corrispondenti dataset esposti. In questo modo, la DL può essere dinamicamente arricchita con nuovi elementi e attributi nel tempo, che possono essere immediatamente utilizzabili online per soddisfare le esigenze degli utenti e migliorare la loro interazione con i dataset. Un altro valore aggiunto di questo approccio risiede nella possibile estrarre dal formato XLSX i metadati per la creazione di percorsi in METS, MODS, EAD o altri standard internazionali, mappando le collezioni con altre biblioteche digitali gestite dagli attuali sistemi DLMS. Infine, le collezioni digitali sono esposte con licenza Creative Commons CC BY SA 4.0<sup>6</sup>, che impone il riconoscimento della paternità all'autore citando la fonte, e la redistribuzione del prodotto derivato dall'alterazione, trasformazione e sviluppo del set di dati per mezzo della stessa licenza attribuita al materiale originario.

La visualizzazione e interrogazione del materiale prodotto. La sezione Dati mostra i collegamenti alle collezioni, e descrive brevemente i fondi originali e il formato CSV e XLSX strutturato per la creazione del percorso dei metadati. Accedendo a un set di dati, l'interfaccia mostra il link per accedere alle collezioni digitali e il relativo percorso dei metadati, nonché ulteriori informazioni, descrivendo una collezione come un'intera risorsa, permettendo agli utenti di visualizzare i record digitali e interagire con essi mediante Preview (Figura 2) e/o effettuando il Download dei file insieme alle descrizioni. Nello specifico, la funzione Preview mostra le stringhe contenenti i metadati valorizzati di ciascuna entità culturale delle collezioni e consente l'accesso alle risorse grazie a strumenti di interrogazione. Queste, attraverso filtri corrispondenti alle etichette del percorso dei metadati, vengono effettuate sia in termini di contenuto che in un'ottica di ricerca dell'oggetto

<sup>6</sup> Creative Commons, Attribuzione 4.0 Internazionale (CC BY 4.0), <https://creativecommons.org/licenses/by/4.0/deed.it> (ultima consultazione 04.05.2023).

digitale. Inoltre, gli utenti possono scaricare il dataset in formato CSV e XLS, come richiesto dal paradigma Open Data. Infine, ogni dataset è associato a un'interfaccia Activity Flows, che offre agli utenti la possibilità di esplorare la provenienza e il ciclo di vita delle risorse, perché integra metadati che registrano e memorizzano tutte le informazioni relative alla creazione, all'aggiornamento e alla modifica di oggetti digitali. Registrando questo tipo di informazioni, i metadati evolvono in patrimonio digitale, affrontando l'art. 2 delle conclusioni del Consiglio UE del 21 maggio 2014 sul patrimonio culturale come risorsa strategica per un'Europa sostenibile [8].

The screenshot shows a web interface for the 'Libretti Rossi' dataset. At the top, there are navigation links and buttons for 'VAL ALLA RISORSA' and 'DATA API'. Below, there's a URL and a note 'File in formato XLSX'. A search bar and a 'Filters' button are visible. The main part of the interface is a table with 535 records, currently showing records 1 through 22. The table has columns: \_id, Identi..., Identi..., Posizi..., Estre..., Estre..., Ogget..., Descr..., Sogg..., Perso..., and Perso... (repeated). The data rows contain various alphanumeric values and names.

_id	Identi...	Identi...	Posizi...	Estre...	Estre...	Ogget...	Descr...	Sogg...	Perso...	Perso...
1	ACOP...	ACOP...	1	1936-...	Isole ...	March...	Carta ...	Direzi...	Direzi...	Antoni...
2	ACOP...	ACOP...	1	1936.0	Isole ...	Zoll Gi...	Verbal...	Direzi...	Direzi...	Zoll Gi...
3	ACOP...	ACOP...	2	1937-...	Isole ...	Ceriol...	Carta ...	Direzi...	Direzi...	Aldo A...
4	ACOP...	ACOP...	2	1937-...	Isole ...	Cresi ...	Verbal...	Direzi...	Direzi...	Giorgi...
5	ACOP...	ACOP...	2	1937-...	Isole ...	Fonta...	Verbal...	Direzi...	Direzi...	Federi...
6	ACOP...	ACOP...	2	1937-...	Isole ...	Torri F...	Carta ...	Direzi...	Direzi...	Franc...
7	ACOP...	ACOP...	3	1938-...	Isole ...	Baiard...	Verbal...	Direzi...	Direzi...	Raffae...
8	ACOP...	ACOP...	3	1938-...	Isole ...	Sardo...	Verbal...	Direzi...	Direzi...	Virgli...
9	ACOP...	ACOP...	4	1937-...	Isole ...	Verbal...	Verbal...	Direzi...	Direzi...	di con...
10	ACOP...	ACOP...	5	1937-...	Isole ...	Allegr...	verbal...	Direzi...	Direzi...	Alfred...
11	ACOP...	ACOP...	1	1939.0	Isole ...	Adam...	Carta ...	Direzi...	Direzi...	Luca ...
12	ACOP...	ACOP...	1	1939.0	Isole ...	Ambr...	Carta ...	Direzi...	Direzi...	Eman...
13	ACOP...	ACOP...	1	1939.0	Isole ...	Amigo...	Carta ...	Direzi...	Direzi...	Fulvio ...
14	ACOP...	ACOP...	1	1939.0	Isole ...	Anegg...	Carta ...	Direzi...	Direzi...	Corra...
15	ACOP...	ACOP...	1	1939.0	Isole ...	Angio...	Carta ...	Direzi...	Direzi...	Giuse...
16	ACOP...	ACOP...	1	1939.0	Isole ...	Antoni...	Carta ...	Direzi...	Direzi...	Antoni...
17	ACOP...	ACOP...	1	1939.0	Isole ...	Ammi...	Carta ...	Direzi...	Direzi...	Franc...
18	ACOP...	ACOP...	1	1939.0	Isole ...	Attana...	Carta ...	Direzi...	Direzi...	Vince...
19	ACOP...	ACOP...	1	1939.0	Isole ...	[Augui...	Carta ...	Direzi...	Direzi...	Ercole...
20	ACOP...	ACOP...	1	1939.0	Isole ...	Autath...	Carta ...	Direzi...	Direzi...	Giova...
21	ACOP...	ACOP...	1	1939.0	Isole ...	Azzari...	Carta ...	Direzi...	Direzi...	Isidor...
22	ACOP...	ACOP...	1	1939.0	Isole ...	Azzari...	Carta ...	Direzi...	Direzi...	Isidor...

Figura 15. Preview del dataset Libretti Rossi - Archivio Storico della Colonia Penale Isole Tremiti

### 3. CONCLUSIONI

Attualmente, la DL è in corso di implementazione. In primo luogo, i set di dati saranno collegati alle immagini prodotte e provenienti da ciascun progetto di digitalizzazione, con l'idea di creare record digitali sostenibili che rispondano ai criteri FAIR [9] per essere facilmente comprensibili dagli utenti. In secondo luogo, si sta creando un'interfaccia di query intuitiva e interattiva che consente una modalità di ricerca sempre più esaustiva che porta l'utente ad approdare direttamente sul singolo record digitale dal quale, attraverso una serie di metadati che ne descrivono il contenuto, i contesti e le relazioni, è possibile raggiungere altri record digitali, rielaborarli e creare così nuova conoscenza e cultura. In particolar modo, l'interfaccia sarà altamente responsive, ossia la veste grafica risponderà in modo diverso, adattandosi al dispositivo dell'utente rendendo l'implementazione più sostenibile essendo unica per i vari dispositivi e sistemi, e consentirà agli utenti di interrogare le collezioni personalizzando la loro ricerca mediante filtri clusterizzati.

In conclusione, il progetto Open Memory Apulia si propone quale sperimentazione di una DL in modalità open, sviluppata secondo una metodologia volta a concepirla come ecosistema digitale culturale, basato non più su una visione attrattiva della cultura ma considerando ogni entità complessa e diversificata come attivatore di interesse capace di interagire e di produrre nuova conoscenza, ottenendo un maggior coinvolgimento degli utenti e migliorare significativamente l'incontro delle comunità antropiche con la cultura e il patrimonio culturale.

### BIBLIOGRAFIA

[1] Tammaro, Anna Maria. «Che cos'è una biblioteca digitale?». DigItalia, 1 (2005).



- [2] Biagetti, Maria Teresa. *Le biblioteche digitali: Tipologie, funzionalità e modelli di sviluppo*. Milano: Franco Angeli Edizioni, 2019.
- [3] Commissione Europea. Comunicazione della Commissione. *Europa 2020. Una strategia per una crescita intelligente, sostenibile e inclusiva*. Bruxelles, 2010. <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=celex%3A52010DC2020> (ultima consultazione 04.05.2023).
- [4] European Commission. *Cultural institutions and cultural Open Data*. 2018. <https://data.europa.eu/en/publications/datastories/cultural-institutions-and-cultural-open-data> (ultima consultazione 04.05.2023).
- [5] Agosti, Maristella. «Digital Libraries». *Mondo Digitale* 43 (2012). [https://mondodigitale.aicanet.net/2012-3/articoli/01\\_digital\\_libraries.pdf](https://mondodigitale.aicanet.net/2012-3/articoli/01_digital_libraries.pdf) (ultima consultazione 04.05.2023).
- [6] Barbuti, Nicola. «Ripensare i formati, ripensare i metadati: prove “tecniche” di conservazione digitale». *Umanistica Digitale* 3(5) (2019). <https://doi.org/10.6092/issn.2532-8816/9055>. <https://umanisticadigitale.unibo.it/article/view/9055/9232> (ultima consultazione 04.05.2023).
- [7] Weston, Paul Gabriele, e Sardo Lucia. *Metadati*. Roma: AIB, 2017.
- [8] Gazzetta ufficiale dell'Unione europea. *Conclusioni del Consiglio del 21 maggio 2014 sul patrimonio culturale come risorsa strategica per un'Europa sostenibile* (2014/C 183/08), 36 C 183 § (2014). [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014XG0614\(08\)&da=IT](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014XG0614(08)&da=IT) (ultima consultazione 04.05.2023).
- [9] Hermon, Sorin, e Niccolucci Franco. «FAIR Data and Cultural Heritage Special Issue Editorial Note». *Int J Digit Libr* 22, 251–255 (2021). <https://doi.org/10.1007/s00799-021-00309-8>. <https://link.springer.com/article/10.1007/s00799-021-00309-8> (ultima consultazione 04.05.2023).

# Nel laboratorio di *Poeti d'Italia in lingua latina*: lavori in corso e prospettive future

Marco Sartor<sup>1</sup>

<sup>1</sup> Università di Parma, Italia – marco.sartor@unipr.it

## ABSTRACT

Il contributo descrive lo stato del progetto *Poeti d'Italia in lingua latina*, una banca dati di autori italiani che accoglie opere latine in versi composte nel periodo compreso all'incirca fra la nascita di Dante e la metà del Cinquecento. In particolare, si presentano i vantaggi di un recente *major upgrade* (MQDQ-Galaxy) che favorisce lo sviluppo di un ambiente collaborativo (Wiki-MQDQ) per l'inserimento di nuovi testi da parte di studiosi privi di competenze specialistiche nell'ambito dell'informatica umanistica mediante un approccio DSL e un *tool* di conversione dei file dal formato *markdown* a XML/TEI.

## PAROLE CHIAVE

Banche dati; archivi digitali; letteratura latina; intertestualità

## 1. INTRODUZIONE

La banca dati *Poeti d'Italia in lingua latina tra Medioevo e Rinascimento* (<https://poetiditalia.it/>) ha preso avvio nel 1999 «con lo scopo [...] di individuare, censire, raccogliere e digitalizzare i componimenti latini in versi, prodotti in area geografica o ambiente culturale italiano, entro il periodo compreso all'incirca fra la nascita di Dante e la prima metà del Cinquecento» ([1]: 38; [2]). Ideato da Paolo Mastandrea e Manlio Pastore Stocchi, con progetto informatico a cura di Luigi Tassarolo ([1]; [3]; [4]), l'archivio elettronico è stato realizzato con bando PRIN (1999-2001) ed è parte della cosiddetta “galassia MQDQ”, espressione con cui si suole indicare una serie di progetti affini il cui risultato più significativo è costituito dal repertorio *Musisque Deoque* (<https://mqdq.it/>), dedicato alla versificazione di età arcaica, classica e tardo-antica [5].<sup>1</sup> Alla data dell'ultimo aggiornamento (29 aprile 2023), la banca dati comprende 859 opere di 316 autori differenti per un totale di circa mezzo milione di versi e oltre tre milioni di *token* (cf. Fig. 1); i testi sono interrogabili per mezzo di un'agile e avanzata maschera di ricerca (per cui cf. [1]: 41-44) e consultabili secondo un doppio ordinamento alfabetico e cronologico che rende lo strumento utilizzabile sia per finalità di studio che in ambito didattico [6].

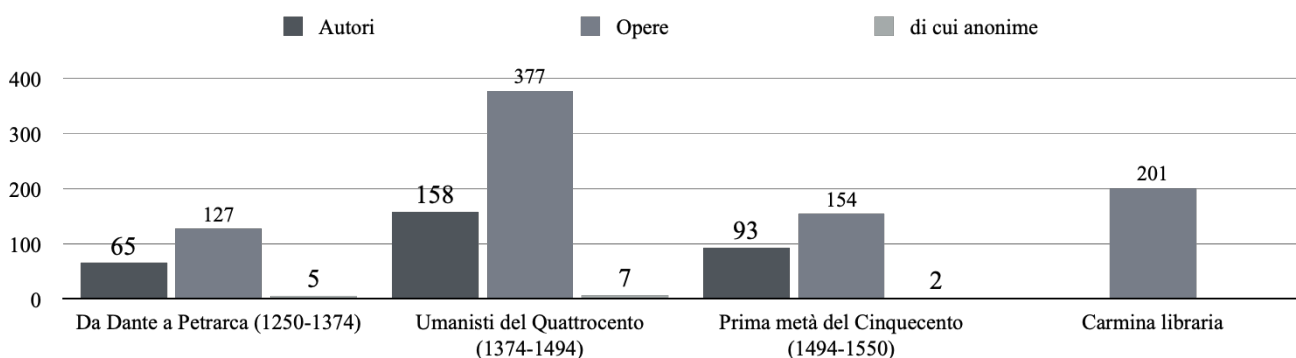


Tabella 1. Alcune statistiche relative ad autori e testi accolti in *Poeti d'Italia*, suddivisi per periodo cronologico.

## 2. L'EDITING COLLABORATIVO

Il mantenimento in linea dell'archivio elettronico per un tempo superiore alle tre decadi è frutto dell'adozione di accorte soluzioni finalizzate a contrastare l'obsolescenza dell'infrastruttura informatica e a favorire l'implementazione del *corpus* anche al termine dei cospicui finanziamenti che ne hanno permesso la genesi. L'ultimo consistente intervento di aggiornamento – ad opera di Luigi Tassarolo e dei ricercatori dell'unità di ricerca veneziana dell'Istituto di Linguistica Computazionale del CNR, con il supporto del centro VeDPH dell'Università Ca' Foscari Venezia – ha favorito lo sviluppo

<sup>1</sup> Alcuni significativi precedenti realizzati dal medesimo gruppo di ricerca sono i repertori su disco *Poesis 2. CD-Rom dei testi della poesia latina* (Bologna, 1999<sup>2</sup>) e *PoetriaNova. A CD-Rom of Latin Medieval Poetry, 650-1250 A. D.* (Firenze, 2010<sup>2</sup>).

di un modello di gestione improntato alla sostenibilità, rendendo l'intera banca dati TEI *compliant*, e ha portato all'allestimento del portale per l'*editing* collaborativo MQDQ-Galaxy (<http://mqdqgalaxy.it>), che costituisce il *back-end* informatico dei progetti della cosiddetta galassia MQDQ (Fig. 1) [7].<sup>2</sup>

Il modello architetturale elaborato nell'ambito dell'implementazione tecnologica è descritto negli atti di questo convegno nel paper *La galassia MQDQ tra innovazione e tradizione* di Angelo Mario Del Grosso, Martina Venuti, Federico Boschetti e Luigi Tessarolo, al quale si rimanda. Questo poster, invece, approfondisce le implicazioni relative a *Poeti d'Italia*, proponendo un *workflow* per la preparazione e l'inserimento di nuove risorse testuali ad opera dei collaboratori. Il punto di partenza è rappresentato dal già citato portale MQDQ-Galaxy, la cui messa a punto accresce sensibilmente la dimensione collaborativa del progetto ([8]: 96-100), comunque già presente al momento dell'ideazione. Imprescindibile *accessus* per ogni studioso che intenda inserire un nuovo testo nella banca dati, lo spazio web contiene diverse linee guida in italiano e in inglese, nonché due *tool* per la conversione dei file delle opere da *markdown* a XML-TEI e viceversa. Lo scopo della documentazione raccolta è quello di consentire agli studiosi di allestire un'edizione digitale indipendentemente dal grado di confidenzialità con i linguaggi di *markup* e, dunque, anche senza il possesso di specifiche competenze informatiche. Per questo, sebbene il testo e i metadati possano sempre essere compilati con una codifica in formato XML-TEI [9], viene offerta la possibilità di utilizzare un approccio orientato ai DSL con la produzione di un file in formato *markdown*. Come si evince dalla Fig. 2, questo si compone in apertura di alcuni metadati espressi nella forma `>name=value`, seguiti del testo dell'opera. Appositi marcatori – anche di carattere metrico – sono posti in corrispondenza delle divisioni (`div`) per segnalare l'inizio di un componimento o di una sezione interna all'opera, ma è possibile inserire elementi descrittivi anche per le eventuali didascalie o per i personaggi, nel caso di opere sceniche. Nella Fig. 2, la notazione `#<1>` (E) indica che il testo è il componimento incipitario della silloge poetica (`<1>`) ed è composto in distici elegiaci (E). La didascalia, riportata alla riga seguente del file, è racchiusa all'interno della coppia di simboli `\[ e ]\`. Quando il file è ultimato, viene caricato sul *tool* di conversione presente in MQDQ-Galaxy: il file XML-TEI prodotto sarà quindi inviato ai gestori del progetto per la revisione e la conseguente pubblicazione online. I documenti testuali che compongono la banca dati sono conservati in *repositories* ad accesso privato, la cui modifica per ragioni di sicurezza è consentita ai soli gestori del progetto; una versione stabile del *corpus* è depositata e accessibile mediante l'infrastruttura CLARIN-IT, dove è identificata da un apposito *handle*.<sup>3</sup>

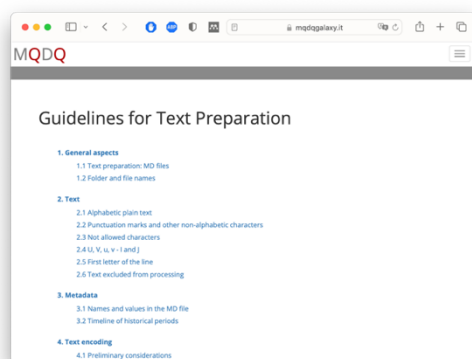


Figura 1. Una schermata del portale MQDQ-Galaxy.

<sup>2</sup> L'accesso al portale è consentito ai soli collaboratori del progetto, che possono richiedere la creazione manuale di un account utilizzando lo strumento di contatto indicato nel *footer* dei siti web della galassia MQDQ.

<sup>3</sup> <https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/OPEN-556>

```

1 -corpuspoetitalia
2 -author-nameTebaldeo, Antonio
3 -author-aliasAntonio Tebaldi detto Tebaldeo
4 -author-pidBenedettoAntonio Tebaldeo
5 -author-abbrevTEBALDEO
6 -author-dateFerrara, 1462/3-Roma, 1537
7 -author-age115
8 -author-number0221_000
9 -work-titleCarmina
10 -work-alttitleCarmina
11 -work-abbrevCar.
12 -work-date1468-1537 ca.
13 -work-formpoetry
14 -work-metervarious
15 -work-number08
16 -bibl-paperS. Pasquazi, 1966
17 -bibl-digitalM. Sartori, I. Tomei, 2022
18
19 <!--(E)
20 [Ad Theophrastum Benedictum Philomusum anicum optimum].
21
22 Ipse tuos solitus continere liber amores,
23 O anime melior pars, Philomuse, mores,
24 Immeritas versu solitus dammare puellas
25 Multaque iactantia uoce superba loqui
26 Vror, ut in mediis ardet seges arida campis
27 Cum furit Aeolii turba sonora dei.
28 Si quis dixerit quendam crudelis "Anabis",
29 Dixissen: "Anticyras i, bibe, stulte, dius".
30 Ipse tamen superatus amo, pemo, uxor, et audax
31 Admissi poena nunc mea lingua luit.
32 Venerat ille dies toto celeberrimus orbe,

```

Figura 2. Un esempio di file in formato *markdown* relativo ai *Carmina* di Tebaldeo

Il flusso di lavoro descritto succintamente nel paragrafo precedente è integrato in un più ampio processo (Fig. 3) che prevede l'impiego di piattaforme come Transkribus<sup>4</sup> (e in futuro eScriptorium<sup>5</sup>) che utilizzano tecnologie OCR e HTR per il riconoscimento e la trascrizione del testo della sorgente (manoscritti od opere a stampa antiche o moderne), nonché di strumenti per l'analisi automatica dei versi latini come Pedecerto<sup>6</sup> per compiere verifiche di carattere metrico su larga scala. Al momento vi sono diversi testi in lavorazione da parte di studiosi che contribuiscono in forma volontaria all'implementazione dell'archivio elettronico: fra questi, Maria Enrica Fedele e Francesca Porro stanno attendendo rispettivamente all'edizione della *Bucolicorum Mimisis* di Curio Lancillotto Pasio (*fl.* XVI sec.) e del *Liber Epistolarum* di Niccolò Bonaiuti (1360 ca.-*post* 1422), entrambi inediti, mentre lo scrivente sta curando l'edizione digitale di alcuni epitaffi e testi in lode di Dante.<sup>7</sup> La reingegnerizzazione dell'architettura tecnologica, la creazione di un ambiente per l'*editing* collaborativo e – non da ultimo – la traduzione in inglese dei contenuti del portale MQDQ-Galaxy testimoniano la volontà di dotare la galassia MQDQ di un *framework* di gestione e sviluppo solido e sostenibile che può essere esportato e impiegato in altri archivi elettronici. Con specifico riferimento a *Poeti d'Italia*, l'auspicio è che nei prossimi mesi il ventaglio dei contributori possa accrescere sensibilmente, accogliendo anche studiosi di istituzioni straniere.

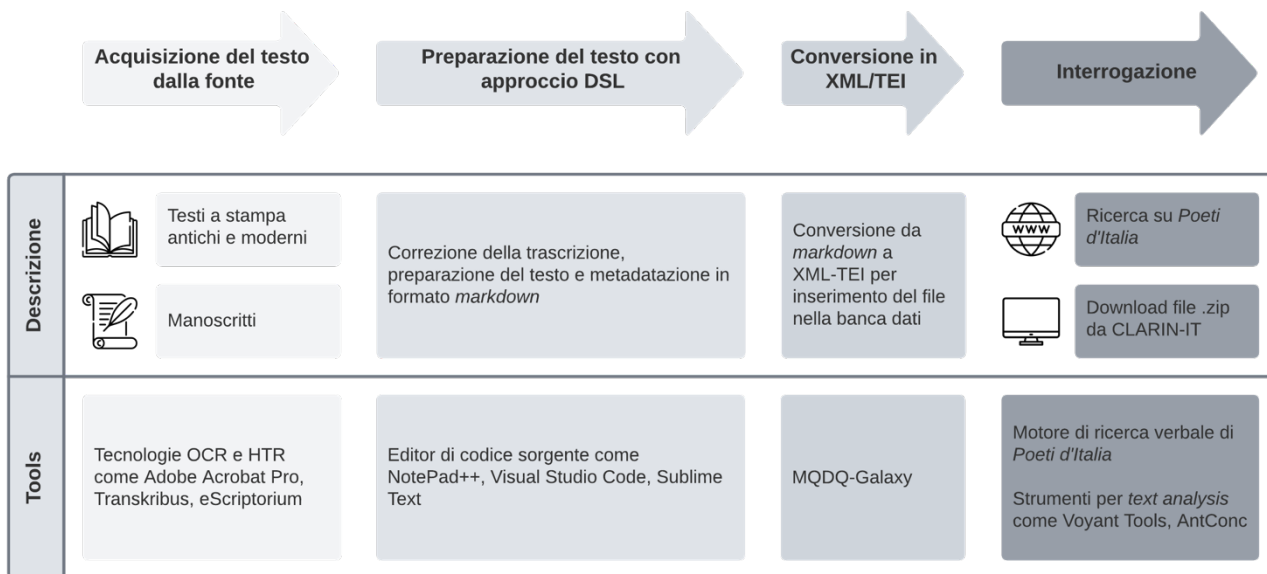


Figura 3. *Workflow* per la preparazione delle risorse testuali.

<sup>4</sup> <http://transkribus.eu>

<sup>5</sup> <https://escriptorium.uk>

<sup>6</sup> <http://pedecerto.eu>

<sup>7</sup> L'elenco di tutti i collaboratori è indicato nell'omonima sezione del sito web: <https://www.poetitalia.it/public/index/collaboratori>. È in corso di allestimento una pagina web che tenga traccia delle più recenti acquisizioni e delle tappe di implementazione del progetto. Gli ultimi testi aggiunti sono i *carmina* di Francesco Pellegrino Ariosto, Malatesta Ariosto, Scipione Balbi, Celio Calcagnini, Girolamo Cantalupi, Ludovico Carbone, Vittoria Colonna, Iacopo Emiliani, Daniele Fini, Alessandro Guarini, Alfonso Guarini, Battista Guarini, Emanuele Guarini, Ludovico Bigo Pittorio, Luca Ripa, Angelo Sabino, Ludovico Sardi, Alberto Tassino, Antonio Tebaldeo.

## BIBLIOGRAFIA

- [1] Mastandrea, Paolo, e Manlio Pastore Stocchi. «Poeti d'Italia in lingua latina. Un archivio elettronico da Dante al primo trentennio del XVI secolo». In *Latino nell'età dell'umanesimo. Atti del Convegno, Mantova, 26-27 ottobre 2001*, 35-50. Accademia Nazionale Virgiliana di Scienze Lettere e Arti 12. Firenze: Olschki, 2004. <https://doi.org/10.1400/48076>
- [2] Mastandrea, Paolo. «Poesia "italiana" in latino: cerca nel web». *Journal of Latin Linguistics* 9, fasc. 3 (2007): 61-70. <https://doi.org/10.1515/joll.2007.9.3.61>
- [3] Mastandrea, Paolo. «Manlio Pastore Stocchi, filologo libero». *Filologia e Critica* XLVI (2021): 241-44.
- [4] Mastandrea, Paolo e Luigi Tassarolo. «Da Musisque Deoque a Memorata Poetis. Le vie della ricerca intertestuale». In *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*, a cura di Maristella Agosti e Francesca Tomasi: 69-77. Padova: CLEUP, 2014.
- [5] Mastandrea, Paolo. «1991-2011, un ventennio di evoluzione nella filologia pratica». In *Convegno di studi in memoria di Giuseppe Gigliozzi. Roma, 27-28 ottobre 2011*, 151-66. Roma: Sapienza Università Editrice, 2012. DOI: <http://dx.doi.org/10.7357/DigiLab-32>
- [6] Zafar, Shahila, e Zaved Ahmed Khan. «Corpus Analysis for Literary Studies: Application and Relevance». In *Learning How to Learn Using Multimedia*, a cura di Deepanjali Mishra e Yuangshan Chuang, 85-93. *Lecture Notes in Educational Technology*. Singapore: Springer, 2021. [https://doi.org/10.1007/978-981-16-1784-3\\_6](https://doi.org/10.1007/978-981-16-1784-3_6)
- [7] Boschetti, Federico, Angelo Mario Del Grosso, e Linda Spinazzè. «La galassia Musisque Deoque: storia e prospettive». In *Paulo maiora canamus. Raccolta di studi per Paolo Mastandrea*, a cura di Massimo Manca e Martina Venuti, 405-19. *Antichistica* 32. Venezia: Edizioni Ca' Foscari, 2021.
- [8] Driscoll, Matthew James, e Elena Pierazzo, a cura di. *Digital Scholarly Editing: Theories and Practices*. Cambridge: Open Book Publishers, 2016. <https://doi.org/10.11647/OBP.0095>
- [9] Burnard, Lou. *What Is the Text Encoding Initiative?. How to Add Intelligent Markup to Digital Resources*. *Encyclopédie Numérique* 3. Marseille: OpenEdition Press, 2014. <https://doi.org/10.4000/books.oep.426>

# Metadata e concetti, una nuova intermediazione delle fonti. Quali prospettive per la ricerca e la diffusione della storia?

Manfredi Scanagatta<sup>1</sup>

<sup>1</sup> Università di Modena e Reggio Emilia, Italia – manfredi.scanagatta@gmail.com

## ABSTRACT

Il mestiere di storici e storiche è destinato a mutare in modo sensibile in conseguenza alla quarta rivoluzione, non solo perché stanno mutando gli strumenti di lavoro, ma perché le fonti, la base su cui si struttura la storia come disciplina, stanno subendo un modificazione ontologica che unita allo sviluppo di sistemi di organizzazione della conoscenza legati all'utilizzo di metadata, richiede di proporre una revisione del loro statuto epistemologico.

Il ruolo che i concetti stanno avendo all'interno dell'organizzazione dell'informazione digitale apre a considerazioni di grande interesse in relazione alla possibilità di costruire una nuova forma di storiografia all'interno di archivi digitali dove il processo di intermediazione avviene in una fase preliminare alla narrazione. Oltre a ciò, in una prospettiva di Public History è di grande interesse la possibilità di costruire oggetti informativi in relazione ad una comunità di riferimento dove i metadata possono essere estrapolati dal contenuto delle fonti.

## PAROLE CHIAVE

Metadata; fonti; storia; storiografia; archivi digitali; Public History; oggetti Informativi

## 1. PROPOSTA

Gli esseri umani sono storicamente esseri informativi producono dati e informazioni da migliaia di anni; ne sono esempi le pitture rupestri, i simboli cuneiformi sumeri, i geroglifici, il sanscrito, l'alfabeto runico, greco e latino; dati che nel loro evolversi, nell'essere contestualizzati e aggregati, hanno generato informazioni, che sono esistite e che esistono grazie all'utilizzo di diverse tecnologie con le quali è stato possibile generare artefatti tangibili.

Il papiro, la pergamena, l'invenzione della carta, l'inchiostro e le penne, la stampa a caratteri mobili, le macchine da scrivere, sono solo alcune delle tecnologie che nel passato hanno consentito la conservazione di un dato in forma di informazione.

Ciò che differenzia la produzione e la condivisione di informazioni tra il passato e il presente è che in passato l'informazione scambiata e registrata aveva una forma tangibile, si presentava a noi come elemento materico, esisteva come oggetto fisico. L'esistenza fisica del dato e dell'informazione ne attestava il valore in quanto documento, nulla di formale esisteva al di fuori di un artefatto documentale fisico. L'essere materico era prova dell'esistenza stessa del documento, che nel suo essere tangibile diveniva parte dell'esperienza umana in società non solo per il contenuto, ma per il fatto stesso che esistesse nel nostro mondo dell'esperienza.

Se i dati, definiti in senso diaforico come elementi puri, protoepistemici<sup>1</sup>, non interpretati, sono rappresentazione dell'agire umano all'interno di un ambiente digitale, e se come teorizza Ferraris<sup>2</sup> la registrazione è alla base della costruzione di un oggetto sociale che deriva da un atto sociale, in che modo storiche e storici possono o potranno definire l'importanza del dato all'interno della loro disciplina?

Questa domanda sorge a partire dalle riflessioni di Bloch in *Apologia della Storia o mestiere di storico*, dove l'autore insiste sulla necessità di osservare e comprendere la storia come disciplina che studia gli uomini in società a partire dalle tracce che hanno prodotto nel loro esistere.

Nella filosofia di Ferraris, specificatamente in *Scienza nuova. Ontologia della trasformazione digitale* possiamo individuare un ragionamento che ben contestualizza all'interno dell'era digitale la necessità di metodo espresse da Bloch.

---

<sup>1</sup> L. Floridi, *The Philosophy of information*, Oxford University Press, Oxford 2011, p. 85

<sup>2</sup> M. Ferraris, *Documentalità. Perché è necessario lasciare tracce*, Laterza, Bari Roma 2019.

L'ipotesi di fondo che mi ha guidato è che il web sia un apparato di registrazione, come la scrittura, il denaro, l'inventario e l'archivio. Il più grande apparato di registrazione che la storia abbia conosciuto fino a qui e il più potente.

D'accordo con la legge fondamentale dell'ontologia sociale, la definizione degli oggetti sociali è oggetto=atto registrato: un oggetto sociale è il risultato di un atto sociale (tale da coinvolgere almeno due soggetti, un soggetto e una macchina delegata, o due macchine delegate) che ha la caratteristica di essere registrato su un qualche supporto<sup>3</sup>.

Se osserviamo il Web come il più grande e potente archivio della storia, dobbiamo chiederci cosa contenga questo archivio. Dati, certamente, ma come definirli? quando sono intellegibili? Quando sono informazione? I dati hanno davvero un valore per una indagine storica che non voglia essere solo quantitativa?

I dati sono alla base dell'esperienza cognitiva umana<sup>4</sup>, esperienza che quando organizzata si manifesta nella strutturazione della società che a sua volta è manifestazione dell'agire umano, ossia di ciò che secondo Bloch deve o dovrebbe interessare lo storico e la storica.

L'oggetto della storia è per natura l'uomo. O meglio: gli uomini. Più che il singolare, favorevole all'astrazione, il plurale, che è grammaticale della relatività, conviene a una scienza del diverso. Dietro i tratti concreti del paesaggio, dietro gli utensili o le macchine, dietro gli scritti che sembrano più freddi e le istituzioni in apparenza più totalmente distaccate da coloro che le hanno fondate, sono gli uomini che la storia vuole afferrare<sup>5</sup>.

Aggiungendo in nota, citando Monod, che cita Michelet al corso all'Ecole normale del 1829 «noi ci occupiamo sia dello studio dell'uomo singolo, e sarà filosofia, sia dello studio dell'uomo sociale, e sarà storia»<sup>6</sup>.

Se è grazie alla tecnologia che l'uomo ha potuto fare il proprio ingresso nella storia<sup>7</sup>, lo storico del futuro dovrà essere in grado di analizzare e criticare fonti prodotte all'interno di società iperstoriche dove gli scambi informativi sono prevalentemente gestiti da sistemi ICT e dove dunque i dati acquistano un valore fondamentale.

Il dato per avere possibilità di divenire informazione ed essere conservato deve passare per un processo di iscrizione, che nell'era pre-digitale mutava a seconda della tecnologia e del medium con il quale si desiderava conservare e rendere fruibile l'informazione.

Il digitale ha uniformato i processi di iscrizione che possono assumere differenti livelli di complessità dati dai linguaggi di programmazione con cui l'umano interagisce con la macchina, ma che sono ontologicamente costituiti da un serie di uno e di zero.

Omologando in senso computazionale i sistemi di iscrizione si ha la possibilità di generare e fruire dell'informazione in modo nuovo. Ciò che era ontologicamente differente ad oggi è composto della stessa sostanza, bit.

La trasformazione digitale che sta coinvolgendo le società contemporanee sta cambiando in modo profondo la nostra relazione con i documenti, ma non di meno – e forse in modo controintuitivo – la scrittura ha ancora un ruolo determinante nella produzione di oggetti sociali.

È necessario interrogarsi sulle implicazioni, i vantaggi e i limiti proposti dai processi di riscrittura (traduzione) dei documenti materici digitalizzati e i processi di sovrascrittura dati dall'utilizzo di paradata e metadati, che consentono di procedere con un ordinamento logico dei documenti in un orizzonte di conservazione. L'utilizzo di paradata e metadati rende possibile la necessaria relazione tra linguaggio umano e linguaggio macchina per rendere i documenti rintracciabili e fruibili.

Ciò che si vuole indagare riguarda le possibilità date dal dato autoconsistente, il Web Semantico, i linguaggi come RIC-CM<sup>8</sup> e i modelli di rappresentazione dell'informazione sviluppati per OAIS<sup>9</sup> attraverso l'utilizzo di standard di metadattazione e vocabolari come DCMI, per immaginare di poter produrre una nuova forma di storiografia dove la metadattazione concettuale delle fonti si mostri come una nuova intermediazione/spiegazione che avviene preliminarmente alla narrazione.

<sup>3</sup> M. Ferraris, *Scienza nuova. Ontologia della trasformazione digitale*, Rosenberg e Sellier, Torino 2018, p. 21.

<sup>4</sup> D. J. Foskett, *Libraries and information systems: a fruitful partnership*. In: *Proceedings of the Annual Conference of CAIS*. Canadian Association for Information Science, 1995, p. 13.

<sup>5</sup> M. Bloch, *Apologia della storia o mestiere di storico*, Einaudi Editore, Torino 1998, p. 22.

<sup>6</sup> *Ibidem* (cfr. nota 4)

<sup>7</sup> L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Raffaello Cortina editore, Milano, 2017

<sup>8</sup> G. Di Marcantonio, F. Valacchi, a cura di, *Descrivere gli archivi al tempo di RIC-CM*, EUM, Macerata 2018.

<sup>9</sup> G. Michetti, *Il modello OAIS*, in «Digitalia», Anno III, n° 1, 2008.



Da una prospettiva di Public History elaborare una rappresentazione dell'informazione – dunque costruire un oggetto informativo – in relazione ad una comunità di riferimento coinvolta in un processo partecipativo di metadating, apre ad interessanti sperimentazioni. Ad esempio come fatto durante il workshop *Metadata*<sup>10</sup>, sviluppato in collaborazione con il Dott. Di Cristofaro durante la Summer School dell'Università di Modena e Reggio Emilia. Utilizzando Omeka S i partecipanti al workshop hanno metadating i documenti estrapolando i contenuti dalle fonti. Il processo di estrazione dei contenuti si è sviluppato su un doppio livello. Al metodo storico di critica e comparazione delle fonti, dal quale si è ottenuto un'analisi semantica utile ad individuare i concetti significativi del documento, si è affiancato un lavoro di analisi linguistica computazionale dei testi, così da restituire risultati di ordine quantitativo e sintattico. Infine si sono comparati i risultati dati dai due processi di analisi – uno qualitativo, l'altro quantitativo – e si è proceduto ad attribuire metadati alla fonte in senso ricorsivo, così da ottenere non solo oggetti digitali, ma oggetti informativi.

Intento del workshop era di estrapolare significati dai documenti per poi riattribuirli al documento in forma di metadato. Senza metadata non esiste oggetto digitale<sup>11</sup>, questa affermazione dovrebbe far porre più di una domanda a storiche e storici che in parte già oggi, ma prevalentemente nel futuro, si dovranno confrontare con fonti digitali.

I metadata non sono solo elementi di contestualizzazione del dato, ma possibili strumenti per costruire significazione concettuale sul documento originale.

Già Weston<sup>12</sup>, sottolineando che i metadati da attribuire a un documento possono essere estrapolati non solo dal contesto ma anche dal contenuto del documento, apre la strada alla possibilità di creare significazione sul documento utilizzando parole e concetti che emergono come rilevanti durante lo studio del documento stesso.

Guaraschi e Guerrini<sup>13</sup> indicano come complessa ma percorribile l'attività di valutazione del messaggio di un documento, cioè del contenuto concettuale di un testo, da utilizzare come strumento di recupero del documento stesso. Anche Tomasi afferma che le Digital Humanities possono valorizzare l'esigenza di lavorare sul contenuto di una risorsa culturale per estrarre nuovi concetti e trasformarli in strumenti di accesso alla conoscenza<sup>14</sup>.

Un lavoro storiografico per prendere forma ha bisogno oltre che dell'analisi critica e comparativa delle fonti, anche della spiegazione e della narrazione che come scrive Topolski sono inevitabilmente guidate anche dal *savoir* dello storico o della storica<sup>15</sup>. Ma se immaginiamo che storici e storiche in seguito all'analisi delle fonti ne estrapolino dei concetti, da trasformare in metadati, questi potrebbero avere il ruolo di *spiegazione*, mentre l'interoperabilità semantica delle fonti, data dall'utilizzo di standard di metadating come DCMI, assumerebbe il ruolo di *narrazione*.

Come scrive Valacchi «alle radici di qualsiasi intento descrittivo sta la comprensione culturale, ci sono cioè le parole con le quali riusciamo a dare spessore alle cose, oltre la nudità dei nomi»<sup>16</sup>.

La narrazione storiografica se sviluppata grazie all'attribuzione di metadata concettuali e alla conseguente interoperabilità delle fonti non avverrebbe attraverso un intreccio per come lo intende White<sup>17</sup>, ma attraverso la spiegazione che è avvenuta grazie ad una concettualizzazione. Il fruitore, guidato dai concetti non seguirà il registro narrativo impostato dallo storico e potrà confrontarsi direttamente con la fonte/oggetto informativo che diviene così elemento intellegibile del discorso storico.

All'interno del Web, nei processi di organizzazione della conoscenza, i concetti assumono un ruolo centrale e mostrano a storici e storiche la possibilità di sviluppare una forma di storiografia che si determina all'interno di archivi digitali e Sistemi Informativi, dove i concetti possono condurre il ricercatore anche non esperto, a confrontarsi con la complessità informativa delle fonti primarie.

Già Koselleck aveva sottolineato l'importanza dei concetti per la spiegazione del contenuto storico

un'elucidazione storica dei concetti via via non debba risalire solo alla storia della lingua, ma anche a dati della storia e della società, è ovvio, poiché ogni semantica, in quanto tale, ha a che fare con contenuti extra linguistici. Questa è la ragione della posizione precaria e marginale che la semantica occupa tra le scienze linguistiche, ma spiega anche perché essa aiuti egregiamente la storia. Attraversando il setaccio della delucidazione concettuale,

<sup>10</sup> <https://www.summerschooldigitalhumanities.unimore.it/2022-edition/scientific-programme/>

<sup>11</sup> P. Feliciati, *Gestione e conservazione di dati e metadati per gli archivi: quali standard?*, Unimc Pubblicazioni Aperte Digitali, 2010 <http://eprints.relis.org/13657/1/atti-testoFeliciati09.pdf>

<sup>12</sup> P.G. Weston, L. Sardo, *Metadati*, Associazione Italiana Biblioteche, Milano 2017

<sup>13</sup> R. Guaraschi, M. Guerrini, *Cos'è l'indicizzazione*, Editrice Bibliografica, Milano, 2022.

<sup>14</sup> F. Tomasi, *Organizzare la conoscenza: Digital Humanities e Web semantico*, Editrice Bibliografica, Milano, 2022.

<sup>15</sup> Jerzy Topolski, *Narrare la storia. Nuovi principi di metodologia storica*, Bruno Mondadori, Milano, 1997 p. 112.

<sup>16</sup> F. Valacchi, *Le parti e il tutto. Integrare la conoscenza*, in «JLIS.it» vol. 13, no. 3 (September 2022).

<sup>17</sup> H. White, *Forme di storia. Dalla realtà alla narrazione*, Carocci Editore, Roma 2006, p. 88.

gli enunciati del passato vengono precisati, le situazioni e i rapporti intesi nel passato diventano più chiari ai nostri occhi nella loro versione linguistica<sup>18</sup>.

Parlando di enunciati del passato non ci si riferisce alla produzione storiografica figlia del lavoro di storiche e storici, ma ai contenuti informativi presenti nelle fonti a noi giunte. Parafrasando Rossana Lista le fonti sono il medium attraverso il quale la realtà storica diventa accessibile<sup>19</sup>.

L'accesso telepistemico alle fonti digitali presente in archivi digitali o sistemi informativi, unita alle possibilità date dall'utilizzo di metadata concettuali, credo ci offra la possibilità di riattualizzare nell'oggi il concetto di *Geschichte* declinato come l'unione tra ciò che è accaduto (*Res Gestae*), la narrazione della storia o dell'avvenimento accaduto (*Historia Rerum Gestarum o Historie*) e la conoscenza degli avvenimenti accaduti, la conoscenza della storia (*Geschichtkunde*)<sup>20</sup>.

A partire da riflessioni di natura metodologica l'intervento proposto si interroga sulla possibilità di utilizzare uno standard come DCMI all'interno di un sistema come Omeka S, per creare significazione sul documento e realizzare oggetti informativi a partire dall'estrapolazione di concetti dalla fonte.

## BIBLIOGRAFIA

- [1] L. Floridi, *The Philosophy of information*, Oxford University Press, Oxford 2011
- [2] M. Ferraris, *Documentalità. Perché è necessario lasciare tracce*, Laterza, Bari Roma 2019
- [3] M. Ferraris, *Scienza nuova. Ontologia della trasformazione digitale*, Rosenberg e Sellier, Torino 2018
- [4] D. J. Foskett, *Libraries and information systems: a fruitful partnership*. In: *Proceedings of the Annual Conference of CAIS*. Canadian Association for Information Science, 1995
- [5] M. Bloch, *Apologia della storia o mestiere di storico*, Einaudi Editore, Torino 1998
- [6] L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Raffaello Cortina editore, Milano
- [7] G. Di Marcantonio, F. Valacchi, a cura di, *Descrivere gli archivi al tempo di RIC-CM, EUM*, Macerata 2018
- [8] G. Michetti, *Il modello OAIS*, in «Digitalia», Anno III, n° 1, 2008
- [9] P. Feliciati, *Gestione e conservazione di dati e metadata per gli archivi: quali standard?*, Unime Pubblicazioni Aperte Digitali, 2010 <http://eprints.rcis.org/13657/1/atti-testoFeliciati09.pdf>
- [10] P.G. Weston, L. Sardo, *Metadata*, Associazione Italiana Biblioteche, Milano 2017
- [11] R. Guaraschi, M. Guerrini, *Cos'è l'indicizzazione*, Editrice Bibliografica, Milano, 2022
- [12] F. Tomasi, *Organizzare la conoscenza: Digital Humanities e Web semantico*, Editrice Bibliografica, Milano, 2022
- [13] Jerzy Topolski, *Narrare la storia. Nuovi principi di metodologia storica*, Bruno Mondadori, Milano, 1997
- [14] F. Valacchi, *Le parti e il tutto. Integrare la conoscenza*, in «JLIS.it» vol. 13, no. 3
- [15] H. White, *Forme di storia. Dalla realtà alla narrazione*, Carocci Editore, Roma 2006
- [16] R. Koselleck, *Futuro passato. Per una semantica dei tempi storici*, Clueb, Bologna 2007
- [17] R. Lista, *Geschichtsbegriff. Storicizzazione del tempo e temporalizzazione della storia*, prefazione a R. Koselleck, *Storia la formazione del concetto moderno*, Clueb, Bologna 2009
- [18] R. Koselleck, *Storia, la formazione del concetto moderno*, Clueb, Bologna 2009

---

<sup>18</sup> R. Koselleck, *Futuro passato. Per una semantica dei tempi storici*, Clueb, Bologna 2007, p. 97.

<sup>19</sup> R. Lista, *Geschichtsbegriff. Storicizzazione del tempo e temporalizzazione della storia*, prefazione a R. Koselleck, *Storia la formazione del concetto moderno*, Clueb, Bologna 2009, p. 14.

<sup>20</sup> R. Koselleck, *Storia, la formazione del concetto moderno*, Clueb, Bologna 2009, p. 40.

# Metodi di armonizzazione per codifiche TEI cooperative: l'edizione Bellini Digital Correspondence

Pietro Sichera<sup>1</sup>, Laura Mazzagufo<sup>2</sup>, Angelo Mario Del Grosso<sup>3</sup>, Daria Spampinato<sup>4</sup>

<sup>1</sup> Ministero dell'Istruzione e del Merito, Italia – [pietro.sichera@posta.istruzione.it](mailto:pietro.sichera@posta.istruzione.it)

<sup>2</sup> Università di Pisa, Italia – [l.mazzagufo@studenti.unipi.it](mailto:l.mazzagufo@studenti.unipi.it)

<sup>3</sup> Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILC), Pisa, Italia – [angelo.delgrosso@ilc.cnr.it](mailto:angelo.delgrosso@ilc.cnr.it)

<sup>4</sup> Istituto di Scienze e Tecnologie della Cognizione (CNR-ISTC), Catania, Italia – [daria.spampinato@cnr.it](mailto:daria.spampinato@cnr.it)

## ABSTRACT\*

Il contributo illustra nel dettaglio tanto le attività quanto gli strumenti inerenti all'armonizzazione dei documenti XML-TEI codificati in maniera cooperativa da studenti di Informatica Umanistica al fine di realizzare l'edizione scientifica digitale delle lettere autografe di Vincenzo Bellini (Bellini Digital Correspondence). Si mostreranno le strategie adottate per automatizzare al massimo la fase di analisi del processo di codifica implementato in seno al progetto. Nel contesto delle attività di supporto all'armonizzazione dei dati è stato sviluppato *NormaTEI*, un tool per l'estrazione delle diverse scelte di codifica sia attraverso il controllo di coerenza strutturale di elementi e gerarchie XML, sia attraverso l'analisi della distribuzione delle frequenze riferite a specifici tag presenti in diverse decine di documenti XML-TEI.

## PAROLE CHIAVE

Cooperative digital scholarly edition; schema-matching; armonizzazione; normaTEI; Bellini Digital Correspondence

## 1. INTRODUZIONE

L'edizione scientifica digitale (DSE, dall'acronimo inglese sciolto in *Digital Scholarly Edition*) "Bellini Digital Correspondence" (BDC)<sup>1</sup> è il prodotto di cinque anni di attività d'indagine scientifica e sviluppo tecnologico sul corpus di lettere autografe del maestro Vincenzo Bellini conservate presso il Museo civico Belliniano di Catania. L'edizione digitale è stata sviluppata a partire da un processo di rappresentazione scientifica del testo veicolato dalla fonte manoscritta adottando le norme XML definite dallo schema di codifica descritto dalle linee guida della Text Encoding Initiative (TEI). In dettaglio, l'edizione image-based è pensata per la fruizione in ambiente web (mediante browser accedendo a internet oppure in locale all'interno del percorso museale) grazie all'utilizzo del software EVT (Edition Visualization Technology) [7]. Il progetto BDC ha avuto una forte valenza didattica: l'attività di codifica delle lettere selezionate è stata svolta anche con l'ausilio degli studenti del corso di Codifica di Testi del Corso di laurea in Informatica Umanistica dell'Università di Pisa, secondo un modello di integrazione tra attività didattica, attività laboratoriale e attività tecnico-scientifica. Le varie fasi del progetto [10] hanno comportato l'analisi, la trascrizione e la codifica del repertorio belliniano, al fine di garantirne un'efficace conservazione, valorizzazione e promozione sia da parte della comunità accademica sia da parte del visitatore occasionale del Museo. Il processo di edizione implementato ha portato all'attenzione del team di lavoro alcune importanti questioni relative alla gestione di un'edizione realizzata in maniera cooperativa quale è stata l'edizione BDC.

L'attività di verifica scientifica e di armonizzazione/allineamento della codifica ha posto alcuni problemi metodologici, dovuti non solo alla varietà delle soluzioni di codifica adottate dagli studenti, ma anche all'eterogeneità dei materiali stessi che costituiscono la fonte primaria. Il corpus epistolare è una collezione di 40 lettere di cui si contano 35 unità codicologiche catturate in digitale da 111 documenti facsimilari: nel computo si annoverano tipologie di risorse molto differenti tra loro quali lettere, minute di lettera e biglietti, ognuno con peculiari caratteristiche formali, strutturali, testuali e materiali, che trovano quindi riscontro in specifiche scelte rappresentazionali. Nel processo di definizione dello schema finale di codifica – e nella conseguente valutazione metodologica –, un caso emblematico è rappresentato da lettere distinte che compaiono nelle stesse unità codicologiche (ad esempio le lettere LL1.23.I e LL1.23.II<sup>2</sup>): la scelta verte tra una codifica orientata alla

---

\* Il presente contributo è stato realizzato dagli autori nella più completa condivisione di impostazioni e di contenuti: nondimeno, la responsabilità del paragrafo 1 va attribuita a D. Spampinato, quella del paragrafo 2 a A.M. Del Grosso, quella del paragrafo 3 a L. Mazzagufo e quella del paragrafo 4 a P. Sichera. Tutti gli autori hanno contribuito in egual misura al paragrafo conclusivo.

<sup>1</sup> <http://bellinidigitalcorrespondence.cnr.it>

<sup>2</sup> La segnatura tiene conto sia dell'aspetto materiale del documento (entrambe le lettere sono infatti identificate come "23"), sia della valenza testuale (nell'eventualità in cui due lettere distinte condividano lo stesso *folio*, queste sono identificate come "I" e "II").

rappresentazione dei dati codicologici e una codifica – adottata dal team di ricerca ma non sempre, in origine, dagli studenti nei propri elaborati – orientata al testo<sup>3</sup>.

L'edizione delle lettere di Bellini ha implementato quindi un modello distribuito (vale a dire in luoghi e tempi diversi), cooperativo (ovvero con diversi soggetti coinvolti nella produzione di molteplici risorse e in differenti attività) e collaborativo (cioè con diversi soggetti coinvolti nella realizzazione della stessa risorsa e condividendo le stesse attività) sviluppando un processo di edizione che ha prodotto non solo la DSE, ma anche riflessioni metodologiche, nonché strumenti digitali finalizzati ad uniformare la rappresentazione dei dati rispetto ad un unico modello rigoroso di testo. Tali strumenti di analisi sono stati utilizzati per selezionare, allineare e correggere ogni singola scelta di codifica: il corpus BDC è quindi così coerentemente codificato in ogni suo fenomeno testuale, paratestuale ed extratestuale. Inoltre, i tool implementati sono del tutto generici e possono essere riutilizzati in altri progetti di DSE. Infatti, l'esperienza del progetto BDC ha sollevato temi di rilevanza metodologica generale per gli studi in Digital Humanities e, in particolare, ha fatto emergere la questione ancora non risolta del confronto e della interoperabilità delle DSE [11]. Il lavoro distribuito di codifica non può evitare totalmente l'eterogeneità delle scelte rappresentazionali per i medesimi fenomeni testuali, sebbene tale eterogeneità sia stata attenuata dai dettami definiti dallo schema di codifica e dal principio di separazione tra rappresentazione e presentazione dell'edizione digitale [4].

In letteratura, la sfida è nota e studiata anche nell'ambito della rappresentazione e della gestione della conoscenza [8]. Non solo esistono strategie di mapping tra schemi formali e categorie concettuali come strumento per allineare la struttura di documenti XML, ma sono stati anche definiti algoritmi detti di *diffing* al fine di rintracciare i luoghi divergenti tra documenti XML, sia per quanto concerne il controllo della struttura sia per il controllo dei contenuti (considerando sia i tag, sia gli attributi) [2, 3, 9].

In ultimo, la comunità scientifica si ripropone di superare alcuni dei problemi evidenziati mediante l'uso dei cosiddetti principi FAIR<sup>4</sup> [5] nel contesto della creazione e pubblicazione di risorse digitali tanto orientate allo studio del testo quanto allo studio della lingua.

## 2. METODOLOGIA

L'edizione del corpus belliniano è frutto di un processo distribuito che si è sviluppato lungo diverse fasi volte alla rappresentazione digitale delle fonti mediante numerose attività svolte in modo collaborativo e cooperativo. In un primo momento, è stato necessario selezionare i materiali a disposizione del Museo, riprodurre la fonte primaria in immagini digitali ad alta risoluzione e raccogliere le informazioni necessarie alla trascrizione dei testi<sup>5</sup>. Successivamente, per agevolare il lavoro degli studenti, è stato creato un modello di codifica istanziato su due lettere belliniane (LL1.4 e LL1.16): per la natura stessa dell'elaborato (soggetto a valutazione in sede d'esame) e la nota flessibilità del vocabolario XML-TEI, i modelli di codifica costituiscono una soluzione aperta e non definitiva per la rappresentazione della fonte primaria. In questa fase, si è provveduto alla stesura dei documenti XML-TEI contenenti le liste dei termini di dominio musicale, delle opere belliniane citate nell'epistolario, dei riferimenti bibliografici e delle entità nominate (persone, luoghi, enti). Tali materiali, insieme alle riproduzioni facsimilari ad alta risoluzione e a una serie di materiali bibliografici, sono stati consegnati agli studenti per avanzare una personale proposta di edizione della lettera assegnata. Gli elaborati – opportunamente caricati su *GitHub* e consistenti in un documento XML contenente la codifica della lettera e vari file a corredo per la visualizzazione web del contenuto testuale – sono stati quindi valutati: è stato necessario correggere o completare la codifica laddove errata o insufficiente; si è dunque proceduto con la validazione formale e scientifica dei documenti codificati, tramite strumenti quali *XMLLint* e *Saxon-HE*. La successiva fase editoriale del progetto si è concentrata sul processo di verifica e armonizzazione delle lettere – necessario per via della molteplicità delle soluzioni di codifica prodotte dagli studenti – e ha previsto:

- la revisione e/o la registrazione di molteplici fenomeni testuali e paratestuali nel documento digitale<sup>6</sup>,
- il completamento delle liste di entità nominate e altri dati notevoli,
- la codifica dei collegamenti a risorse e repertori autorevoli disponibili sul web secondo le pratiche dei *Linked Open Data* (ad esempio, *VIAF*, *GeoNames*, *RISM*),
- la revisione delle regioni d'interesse a partire dalle fonti facsimilari.

<sup>3</sup> Nella pratica, tale scelta si è tradotta nella produzione di un documento XML-TEI per ogni missiva, indipendentemente dalla disposizione delle lettere sul supporto, laddove gli elaborati degli studenti riunivano in un unico file la codifica di eventuali testi 'multipli', riconducibili a missive differenti ma contenuti nella stessa unità codicologica.

<sup>4</sup> FAIR sta per Findable, Accessible, Interoperable, Reusable.

<sup>5</sup> Per la trascrizione dei testi delle lettere, si è fatto riferimento alla più recente e autorevole edizione a cura di Graziella Seminara [1].

<sup>6</sup> Il processo di edizione scientifica digitale è analizzato nel dettaglio in [6].

### 3. ARMONIZZAZIONE DELLA CODIFICA

Il lavoro di armonizzazione si è concentrato in particolare su due aspetti diversi: quello grafico e quello testuale.

#### 1. Armonizzazione grafica

L'armonizzazione grafica si è concentrata sulla corretta individuazione e metadateazione delle zone di interesse, correggendo o integrando quelle individuate precedentemente. Questa attività ha portato alla realizzazione di un software per la codifica delle zone di interesse, chiamato *ZoneRW*<sup>7</sup>.

*ZoneRW* è uno strumento interattivo dotato di una intuitiva interfaccia grafica che, tramite la gestione delle varie istanze dell'elemento <surface> e dei relativi nodi figlio <graphic> e <zone>, non solo mostra le regioni di interesse registrate nel repertorio digitale facendo uso dei propri attributi @xml:id e @rendition, ma ne permette la modifica estemporanea mediante l'aggiunta o la rimozione delle coordinate delle zone, generando di conseguenza il corrispondente codice XML-TEI.

#### 2. Armonizzazione testuale

L'armonizzazione testuale si è rivelata da subito un'attività assai complessa. Infatti, la possibilità – concessa dalle linee guida TEI – di rappresentare i medesimi fenomeni in modi diversi, benché egualmente validi, pone l'editore di fronte a un'ampia varietà di possibili soluzioni, difficilmente superabile in fase di definizione dello schema.

La codifica di una risorsa testuale in formato XML-TEI è quindi per natura non univoca, bensì interpretabile e soggetta a sfumature semantiche derivanti dalle intenzioni del curatore digitale. Si pensi, a titolo di esempio, a come il riferimento a una persona o al suo nome possa essere codificato tanto con <rs type="person"> quanto con <name type="person">, oppure con <persName>. Le medesime considerazioni possono essere avanzate per molti altri fenomeni, tra cui la punteggiatura (codificata o meno), la presenza delle note (nel back o inline), i riferimenti generici (<ref> o <ptr>). Oltre a ciò, nel proprio elaborato non sempre gli studenti hanno seguito con rigore la struttura proposta dal modello di codifica e dalla documentazione operativa prodotta: è stata spesso attestata, infatti, una codifica sintatticamente corretta ma non conforme al modello oppure eccessivo rispetto alle necessità e alle finalità del progetto (e.g. l'aggiunta di elementi funzionali alla visualizzazione web ideata dagli studenti).

Per uniformare la codifica si sarebbe potuto ricorrere a un processo completamente manuale; è parso però chiaro fin da subito che tale procedura non sarebbe stata risolutiva per i seguenti motivi:

- molteplicità dei revisori e conseguente possibilità di geminazione delle soluzioni correttive;
- presenza di gerarchie XML non corrispondenti, benché simili e soddisfacenti i requisiti della validazione;
- presenza di refusi ed errori semantici.

Appare chiaro dunque come le codifiche dovessero essere uniformate sia per garantire omogeneità al lavoro, sia per consentire un'efficace trasformazione dell'edizione in un documento TEI conforme, compatibile con il software di visualizzazione EVT. A tal proposito, il ricorso ai già citati algoritmi di *diffing* XML<sup>8</sup> non risulta tuttavia applicabile, almeno per la fase di armonizzazione, per varie ragioni:

- contrariamente al consueto processo di *diffing*, che riguarda due o più versioni di file simili, nel progetto BDC i file da confrontare, descrivendo unità testuali differenti, erano molto dissimili tra di loro;
- erano inoltre presenti valori identici per differenti percorsi nell'albero XML (*path*): lo stesso valore si può ritrovare correttamente su tag con *path* differenti;
- in aggiunta, erano correttamente assegnati valori identici per tag differenti, stessi valori per attributi identici su tag differenti o valori identici indipendentemente dall'attributo.

### 4. IL TOOL *NORMATEI* PER L'ARMONIZZAZIONE IN XML-TEI

Per supportare la fase di armonizzazione, il team di BDC ha quindi ritenuto opportuno sviluppare un tool per il controllo dei fenomeni codificati in XML-TEI in maniera complessiva ed organica per tutte le risorse in esame. L'implementazione dell'algoritmo ha portato allo sviluppo di un componente che può essere integrato nello strumento *ZoneRW* oppure eseguito come strumento *stand-alone*.

<sup>7</sup> <https://doi.org/10.5281/zenodo.5599510>

<sup>8</sup> L'osservazione riguarda sia gli algoritmi di controllo strutturale sia gli algoritmi di controllo contenutistico, in base alla distinzione effettuata, tra gli altri, da Cuculovic [3].

Il tool *NormaTEP*<sup>9</sup> ha innanzitutto il compito di raccogliere complessivamente i dati di codifica presenti nei file. L'algoritmo implementato (*Figura 1*) restituisce il nome degli elementi, degli attributi e dei relativi valori.

Per ognuno dei file XML-TEI da analizzare, l'algoritmo crea un puntatore alla radice della struttura XML del file (il tag <TEI>) e lancia la procedura *ScanSonsAndSiblings*. Questa prende come parametro il puntatore a un nodo (alla prima esecuzione è la radice della struttura del file) e viene chiamata ricorsivamente per ogni figlio del nodo; al termine di ogni ricorsione vengono memorizzati il tag attuale e il suo valore, il suo percorso completo, i suoi attributi e i rispettivi valori. In questo modo si ottengono tutti i dati e le informazioni per l'intero albero XML-TEI di ogni file. Dall'osservazione di questi dati è possibile eseguire una serie di controlli sulla similarità. A titolo esemplificativo possiamo citare i seguenti controlli:

- più tag con il medesimo percorso potrebbero dover assumere lo stesso valore;
- tag con lo stesso nome e con lo stesso livello di profondità o simile (tutt'al più con una oscillazione di un livello nella gerarchia) potrebbero dover avere lo stesso percorso;
- più attributi con lo stesso nome potrebbero dover registrare lo stesso valore;
- attributi identici riferiti allo stesso tag potrebbero dover assumere lo stesso valore;
- più tag con eguale gerarchia potrebbero dover avere gli stessi attributi;
- valori di attributo identici potrebbero dover avere lo stesso nome di attributo.

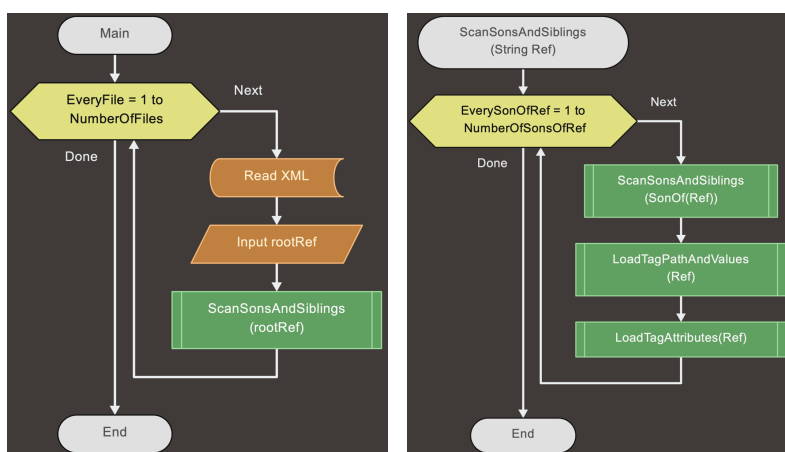


Figura 1. Diagramma di flusso per l'algoritmo implementato all'interno del tool di supporto all'armonizzazione *NormaTEI*.

Il tool possiede una sua specifica interfaccia di ricerca (*Figura 2*) e consente l'esportazione dei dati (tag, percorso, profondità, attributi, valori) in fogli di calcolo Excel.

Prospetto tag con attributi

401 selezionati su 23.567

File	Percorso	note	Tag	ValoreTag	Attributo	ValoreAttributo
LL1_33.xml	/TEI/teiHeader/file...	note		La data di regrad...		
LL1_34.1.xml	/TEI/teiHeader/file...	note		Si ignora se la min...		
LL1_35.xml	/TEI/teiHeader/file...	note		Il nome del destin...		
LL1_1.xml	/TEI/teiHeader/file...	handNote			xml:id	h3
LL1_1.xml	/TEI/teiHeader/file...	handNote				
LL1_1.xml	/TEI/teiHeader/file...	handNote			xml:id	h2
LL1_1.xml	/TEI/teiHeader/file...	handNote			facs	#LL1.1_hotspot_1r_01
40	6	2	40	7	132	

Figura 2. Interfaccia di ricerca del tool *NormaTEI*.

<sup>9</sup> <https://doi.org/10.5281/zenodo.7703053>

L'utilizzo del tool *NormaTEI* all'interno del progetto BDC ha posto in evidenza il sostanziale dimezzamento del numero dei differenti percorsi XML-TEI (1148) annotati nell'iniziale codifica degli elaborati, rispetto al conteggio effettuato dopo la fase di armonizzazione (594).

In aggiunta allo specifico strumento di ricerca, vengono messi a disposizione ulteriori strumenti di natura statistica che consentono sia un'analisi della distribuzione dei fenomeni testuali sia un controllo sulla consistenza dei dati.

I tool generano una struttura *a frame* costituita da molteplici colonne (*Figura 3*): la prima colonna contiene un singolo tag oppure un percorso completo e la seconda colonna mostra il numero di volte in cui quel fenomeno occorre. A partire dalla terza colonna vengono elencati il nome del documento in cui il tag/percorso non è presente (quando il fenomeno è presente in più della metà dei file) oppure i documenti in cui quel tag/percorso è presente (quando il fenomeno è presente in meno della metà dei file).

A	B
TAG	QUANTI FILE QUALI FILE (presente o mancante)
/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/handDesc/handNote/	39 LL1_5.xml
/TEI/text/body/div/ab/s/	38 LL1_31.I.xml LL1_31.II.xml
/TEI/text/body/div/ab/s/lb/	38 LL1_31.I.xml LL1_31.II.xml
/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/handDesc/handNote/p/	37 LL1_1.xml LL1_2.xml LL1_5.xml
/TEI/text/back/div/div/	36 LL1_12.xml LL1_13.I.xml LL1_19.xml LL1_34.I.xml
/TEI/text/body/div/salute/	36 LL1_9.xml LL1_14.xml LL1_31.I.xml LL1_31.II.xml
/TEI/text/body/div/salute/s/	35 LL1_9.xml LL1_14.xml LL1_21.xml LL1_31.I.xml LL1_31.II.xml
/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/note/	34 LL1_6.xml LL1_11.xml LL1_13.II.xml LL1_16.xml LL1_22.xml LL1_30.xml
/TEI/text/body/div/ab/fw/	34 LL1_1.xml LL1_3.xml LL1_5.xml LL1_7.xml LL1_15.xml LL1_19.xml
/TEI/text/body/div/salute/s/lb/	34 LL1_9.xml LL1_14.xml LL1_21.xml LL1_25.xml LL1_31.I.xml LL1_31.II.xml
/TEI/text/body/div/ab/fw/gap/	30 LL1_1.xml LL1_3.xml LL1_5.xml LL1_6.xml LL1_7.xml LL1_8.xml LL1_12.xml LL1_15.xml
/TEI/text/back/div/div/note/	27 LL1_7.xml LL1_8.xml LL1_12.xml LL1_13.I.xml LL1_19.xml LL1_21.xml LL1_22.xml LL1_27
/TEI/text/back/div/div/note/p/	27 LL1_7.xml LL1_8.xml LL1_12.xml LL1_13.I.xml LL1_19.xml LL1_21.xml LL1_22.xml LL1_27
/TEI/text/body/div/ab/fw/idno/	27 LL1_1.xml LL1_3.xml LL1_5.xml LL1_7.xml LL1_9.xml LL1_11.xml LL1_13.I.xml LL1_15
/TEI/text/body/div/ab/s/persName/	27 LL1_12.xml LL1_13.I.xml LL1_13.II.xml LL1_17.xml LL1_19.xml LL1_20.xml LL1_21.xml
/TEI/text/body/div/ab/s/rs/	27 LL1_12.xml LL1_13.I.xml LL1_16.xml LL1_21.xml LL1_23.I.xml LL1_23.II.xml LL1_24.II
/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/note/persName/	26 LL1_1.xml LL1_6.xml LL1_9.xml LL1_11.xml LL1_13.I.xml LL1_13.II.xml LL1_16.xml LL1
/TEI/text/body/div/ab/s/choice/	26 LL1_12.xml LL1_13.I.xml LL1_16.xml LL1_17.xml LL1_19.xml LL1_21.xml LL1_23.I.xml
/TEI/text/body/div/ab/s/w/	26 LL1_12.xml LL1_18.xml LL1_19.xml LL1_21.xml LL1_22.xml LL1_23.I.xml LL1_23.II.xml
/TEI/text/body/div/ab/s/choice/abpr/	25 LL1_12.xml LL1_13.I.xml LL1_16.xml LL1_17.xml LL1_19.xml LL1_21.xml LL1_23.I.xml

Figura 3. Struttura a frame dei dati estratti mediante l'uso di *NormaTEI* visualizzati.

L'utilità dei tool risiede in due aspetti legati al tipo di *output*: 1) è possibile conoscere in quali documenti un fenomeno non è presente per controllare e analizzare il motivo della mancata presenza; 2) è possibile registrare i fenomeni al fine di controllare e analizzare il motivo della bassa frequenza di attestazione. In entrambi i casi il tool permette quindi di trovare errori e particolarità della codifica, oppure di controllare il motivo per cui un fenomeno è presente su gran parte del corpus ma non sul corpus completo.

Infine, come mostrato in *Figura 4*, sebbene gli algoritmi di *diffing* non siano stati utilizzati per allineare i documenti XML ed estrarre le statistiche di struttura e codifica, sono stati tuttavia impiegati (tramite GitHub) per tenere traccia dello stato di avanzamento del processo di armonizzazione, controllando i luoghi divergenti ed evidenziando le modifiche apportate alla struttura e al contenuto dei dati. Si notano in figura le modifiche al nome degli elementi, al nome degli attributi e al valore del contenuto sia degli elementi e sia degli attributi.

erait	214	erait
</sic>	→ 215+	</del>
<corr resp="TEI-ListPerson.xml#VB" rend="above">	216+	<add place="above">
était	217	était
</corr>	→ 218+	</add>
</choice>	219+	</subst>
<gap xml:id="LL1_31_hs_2f_05" facs="#LL1_31.I_hotspot_2f_05"	220+	<gap agent="ink" reason="illegible" extent="one-word" />
reason="ink blot" agent="ink" quantity="1" unit="word"/>	221+	</s>

Figura 4. *Diffing* tra due diversi *commit* di *git* all'inizio e durante il processo di armonizzazione per la lettera n. LL1.31.

## 5. CONCLUSIONI

Quanto descritto nel contributo dimostra come spesso i progetti di ricerca scientifica portino a riflessioni e a risultati inaspettati. Se l'impatto didattico del processo distribuito di codifica era immaginabile nei confronti degli studenti, così come l'intrinseca valenza scientifica del progetto, certamente inaspettato è stato il delinarsi di una metodologia volta ad affrontare lavori di codifica condivisi. Ciò ha portato allo sviluppo di soluzioni innovative per risolvere problemi ricorrenti mediante soluzioni generiche. Tali esiti non solo si applicano alla realizzazione di edizioni collaborative XML-TEI, ma più in generale, a qualsiasi attività di codifica condotta con linguaggi di marcatura descrittivi.

## 6. RINGRAZIAMENTI

Il progetto BDC è stato in parte finanziato dal Patto per Catania a valere sul Fondo Sviluppo e Coesione 2014-2020. Gli autori, inoltre, ringraziano gli studenti dell'insegnamento di Codifica di Testi del corso di laurea in Informatica Umanistica



(Università di Pisa) e in particolare Carla Congiu e Santa Pellino, per il lavoro svolto nell'ambito dei propri progetti di tesi di laurea.

## BIBLIOGRAFIA

- [1] Bellini, V. Carteggio. Edited by Graziella Seminara. Firenze: Olschki, 2017.
- [2] Ciancarini, P., A. Di Iorio, C. Marchetti, M. Schirinzi, e F. Vitali. "Bridging the Gap between Tracking and Detecting Changes in XML." *Software: Practice and Experience* 46, n. 2 (2016): 227-50. <https://doi.org/10.1002/spe.2305>
- [3] Cuculovic M., F. Fondement, M. Devanne, J. Weber, e M. Hassenforder. "Semantics to the rescue of document-based XML diff: A JATS case study." *Software: Practice and Experience* 52, n. 6 (2022): 1496-516. <https://doi.org/10.1002/spe.3074>
- [4] Cummings, J. "Opening the Book: Data Models and Distractions in Digital Scholarly Editing." *International Journal of Digital Humanities* 1 (2019): 179-93. <https://doi.org/10.1007/s42803-019-00016-6>
- [5] De Jong, F., B. Maegaard, K. De Smedt, D. Fišer, e D. Van Uytvanck. "CLARIN: Towards FAIR and Responsible Data Science Using Language Resources." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*: 3259-64. <https://aclanthology.org/L18-1515>
- [6] Del Grosso, A. M., E. Capizzi, S. Cristofaro, M.R. De Luca, E. Giovannetti, S. Marchi, G. Seminara, D. Spampinato. Bellini's Correspondence: a Digital Scholarly Edition for a Multimedia Museum. *Umanistica Digitale* 3(7): 23-47, 2019. DOI: <http://doi.org/10.6092/issn.2532-8816/9162>
- [7] Di Pietro, C., e R. Rosselli Del Turco. "Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions." *Digital Scholarly Editions as Interfaces* (2018): 133-63.
- [8] Euzenat, J., e P. Shvaiko. "The Matching Problem." *Ontology Matching* (2013): 25-54. [https://doi.org/10.1007/978-3-642-38721-0\\_2](https://doi.org/10.1007/978-3-642-38721-0_2)
- [9] Jang, B., S. Park, e Y. Ha. "A Stream-Based Method to Detect Differences between XML Documents." *Journal of Information Science* 43 n. 1 (2017): 39-53. <https://doi.org/10.1177/0165551515602805>
- [10] Pellino, S., P. Sichera, A.M. Del Grosso, D. Spampinato. "Dalla codifica alla fruizione: l'edizione digitale Bellini Digital Correspondence". In F. Ciraci, G. Miglietta, C. Gatto (Eds.), *AIUCD 2022 - Proceedings. Quaderni di Umanistica Digitale* (2022): 163-168.
- [11] Schmidt, D. "Towards an Interoperable Digital Scholarly Edition." *Journal of the Text Encoding Initiative* 7 (2014). <https://doi.org/10.4000/jtei.979>

# LINEA: per un portale della letteratura italiana del Novecento

Antonio Sicchera<sup>1</sup>, Liborio Pietro Barbarino<sup>2</sup>, Christian D'Agata<sup>3</sup>, Giuseppe Palazzolo<sup>4</sup>

<sup>1</sup> Università degli Studi di Catania, Italia – asicchera@unict.it

<sup>2</sup> Università degli Studi di Catania, Italia – liborio.barbarino@unict.it

<sup>3</sup> Università degli Studi di Catania, Italia – christian.dagata@gmail.com

<sup>4</sup> Università degli Studi di Catania, Italia – giuseppe.palazzolo@unict.it

## ABSTRACT

Il contributo presenta il progetto LINEA, prototipo di un'edizione-archivio della letteratura italiana del Novecento che integra i diversi aspetti orbitanti intorno al testo letterario, tra filologia, lessicografia, ermeneutica, didattica e dimensione *public*. Dalle edizioni facsimile, sinottiche e intertestuali di *Lavorare stanca* e *Il nome della rosa* fino al vocabolario de *La luna e i falò*, passando per i contenuti didattici di «FuoriClasse», vengono presentati e discussi alcuni modelli di uso del digitale per la valorizzazione del Novecento letterario.

## PAROLE CHIAVE

Letteratura italiana contemporanea; Digital Scholarly Edition; lessicografia; Cesare Pavese; Umberto Eco

## 1. INTRODUZIONE. IL PROGETTO

Il progetto LINEA<sup>1</sup> (Letteratura Italiana del Novecento: per un'Edizione-Archivio), nato all'interno del Piano di ricerca "Prometeo" promosso dall'Ateneo di Catania, ruota attorno alla creazione di un prototipo di edizione-archivio (*Figura 1*) della letteratura italiana del Novecento, concepito all'interno di un ecosistema digitale nel quale possano convivere ed alimentarsi a vicenda la realizzazione di risorse scientifiche digitali e di contenuti *public* con una curvatura verso la didattica della letteratura del Novecento.

Obiettivo primario è dunque l'elaborazione di edizioni scientifiche digitali filologicamente accurate [9] di 'campioni' significativi della prosa e della poesia del Novecento che possano sopportare molteplici esperienze di lettura. La critica e l'ermeneutica letteraria dialogano dunque con gli strumenti d'indagine lessicografica e semantica, anche in una dimensione comparatistica, mantenendo sullo sfondo la prospettiva didattica e divulgativa. Obiettivi specifici sono la realizzazione di una banca dati interrogabile a più livelli attraverso un'interfaccia web, che integri un database SQL e un web server, affiancata dai testi codificati in XML<sup>2</sup>-TEI<sup>3</sup> [15] e visualizzati con EVT<sup>4</sup> [11], in modo da favorire l'interoperabilità in linea coi principi FAIR<sup>5</sup>. Il portale è dunque concepito come un sistema integrato di dati centrati sul testo letterario, ma organizzati secondo una rete che include anche apparati critici, varianti d'autore, vocabolari, materiali preparatori, percorsi didattici e altre espansioni.

Le metodologie adoperate sono quelle che ad oggi hanno offerto maggiori garanzie di tenuta e affidabilità: a) codifica, strutturazione e trattamento informatico dei testi letterari; b) lessicografia letteraria e statistica lessicale; c) data visualization. La metodologia concordanziale, sottesa a tutto il progetto, rappresenta infatti un'innovazione nella tradizione: utilizza un metodo tipico degli studi antichi e medievali, volto a una penetrazione puntuale dei testi (la Bibbia in primis), integrandolo con le tecnologie digitali dell'era contemporanea, al fine di farne emergere le potenzialità critiche e interdisciplinari. Il sistema di elaborazione testuale, archiviazione, generazione e presentazione dei dati del progetto – che si avvale della lezione del pioniere degli studi lessicografici assistiti dal computer, Roberto Busa SJ<sup>6</sup> – era stato ideato già negli anni Ottanta da Giuseppe Savoca [13]. Esso si pone a fianco, con caratteristiche del tutto peculiari ed originali, degli archivi testuali prodotti successivamente: sia offline – come la LIZ (1993 e 2001) [14] – che online, come il monumentale corpus TLIO<sup>7</sup>, realizzato dall'OVI.

---

<sup>1</sup> <https://linea.unict.it/>

<sup>2</sup> <https://www.w3.org/XML/>

<sup>3</sup> <https://tei-c.org/>

<sup>4</sup> <http://evt.labcd.unipi.it/>

<sup>5</sup> <https://www.go-fair.org/fair-principles/>

<sup>6</sup> <https://www.corpusthomicum.org/it/index.age>

<sup>7</sup> <http://tlio.oivi.cnr.it/TLIO/>

Va infine menzionato il contesto in cui lo sviluppo del progetto si colloca: l'équipe del Centro di Informatica Umanistica (CINUM) è attivamente coinvolta nell'Edizione Nazionale dell'Opera Omnia di Luigi Pirandello, di cui cura il portale PirandelloNazionale [4], e nella realizzazione di un prototipo di archivio digitale della poesia italiana dell'Otto/Novecento (ALPION) al fine di rendere disponibile alla comunità scientifica le concordanze realizzate da Savoca con il sistema IBM AS/400 [4]. In particolare, dal portale PirandelloNazionale il progetto LINEA eredita la concezione aperta e dinamica, l'integrazione di risorse di natura diversa, le molteplici possibilità di studio e di lettura. LINEA mette quindi in dialogo la metodologia concordanziale con gli attuali sistemi di digitalizzazione e codifica dei testi, allestendo dei modelli di edizione digitale conformi allo standard *de facto* della TEI.

### Che cos'è LINEA?

La linea della grande letteratura è una linea spezzata, controversa, mai affidata a segni netti, incontrovertibili, banali. Eppure il suo nitore e la sua pulizia emergono dalle fratture della vita e dal fango della storia per ridare senso e luce, come se si creasse improvvisamente uno spazio bianco e potessimo improvvisamente seguirne le mosse. Dare corpo e conferire durata a queste grandi parole è il compito di Linea, edizione e archivio insieme dei testi maggiori della letteratura italiana del Novecento. Un archivio in progress, non un museo, un'edizione viva e non una trafia di stanchi apparati, una parola per molti e non il patrimonio di pochi specialisti.



CESARE PAVESE

LAVORARE STANCA

*Il primo libro di Pavese, in due edizioni (Solaria, Firenze 1936 ed Einaudi, Torino 1943) diverse per forma e contenuto. Un esempio tra i più luminosi di poesia impura, a lungo ritenuto da chi l'ha scritto, in terza persona, «il suo libro più significativo».*



CESARE PAVESE

LA LUNA E I FALÒ

*La luna e i falò è l'ultimo romanzo di Cesare Pavese. Pubblicato nel 1950, racconta del ritorno di Anguilla al paese nel quale è cresciuto. L'opera condensa in sé i temi e i motivi fondanti della scrittura pavesiana, ed è considerata la summa della sua produzione.*



UMBERTO ECO

IL NOME DELLA ROSA

*L'opera prima di Umberto Eco. Di volta in volta definita bestseller di qualità, conte philosophique, giallo storico, capolavoro del postmoderno. Una riflessione sul riso e sulla verità che ha dato vita a innumerevoli adattamenti, rifacimenti, traduzioni.*

Figura 1. Home del portale LINEA.

## 2. EDIZIONI E VOCABOLARI

L'edizione Archivio nasce dalla necessità di indagare e rappresentare le diverse possibilità della Letteratura del Novecento. La sua natura di 'archivio' consente di mettere insieme testi diversi, edizioni differenti, materiali di natura eterogenea, accomunati da un obiettivo comune: dare al lettore la possibilità di leggere il testo in modo molteplice, multimediale, interattivo. In particolare, sono state progettate diverse tipologie di edizioni:

- a. sinottiche (o edizioni a confronto), che permettono al lettore di leggere le diverse redazioni di un testo in modo dinamico (confrontando tra loro edizioni a stampa o edizioni e autografi): interagendo con il testo è possibile scegliere i testimoni e confrontare le diverse lezioni. I tag usati sono quelli del *Parallel segmentation method* delle *Guidelines* TEI, marcando l'entrata di apparato (<app>) e offrendo la lezione principale a testo (<lem>) e le varianti (<rdg>). L'interfaccia scelta è quella di EVT2;
- b. facsimile, ovvero dei veri e propri iconotesti che permettono al lettore di entrare direttamente nel laboratorio dello scrittore, confrontandosi con gli autografi e facendo esperienza del testo come «approssimazione al valore» ([2]). Tra i tag usati <facsimile> e <surface> elaborati con il *tool* TeiZoner<sup>8</sup>. Si propone come interfaccia EVT1;
- c. intertestuali, pensate per permettere di entrare nel laboratorio dello scrittore da una porta diversa rispetto a quella dell'edizione facsimile: l'intertestualità come «dialogo fra più scritture» [8] e «palinsesto» [6]. I tag usati sono <quote> per le fonti e <seg> per i passi paralleli;

<sup>8</sup> <http://teicat.huma-num.fr/zoner.php>

- d. apparato, che presentano soltanto le varianti con un minimo di contesto (brevi estratti del testo), dal momento che lavorare col Novecento significa inevitabilmente confrontarsi con testi tutelati dal diritto d'autore. I tag usati sono quelli del *Parallel segmentation method* ed è una variante dell'edizione sinottica;
- e. *digitized* e *born digital*, edizioni critiche 'tradizionali', nate per il digitale, ma che ugualmente rientrano tra le edizioni *digitized* [12], in quanto riprendono la forma tradizionale dell'edizione critica cartacea e possono essere agevolmente scaricate e fruite con qualunque *device*.

Partendo da questa prospettiva sono stati realizzati e proposti i lavori su Cesare Pavese (*alpha e omega: Lavorare stanca, La luna e i falò*) e Umberto Eco (*Il nome della rosa*, il suo libro certamente più noto). L'edizione sinottica (Figura 2) presenta quindi le due edizioni di *Lavorare stanca* (Firenze 1936 e Torino 1943) [4], che divergono nettamente nell'architettura testuale e nella portata semantica complessiva. Tale sproporzione può ridurre sottotraccia o peggio sotto silenzio il livello microtestuale, che, nei componimenti in comune alle due stampe, presenta poche varianti ma non tutte trascurabili. L'edizione digitale, di semplice interrogazione, consente di raffrontare sinotticamente tali varianti e afferrare i segni visibili di una ricerca poetica che muove su altre sfere. D'altra parte, l'edizione facsimile (Figura 3) della poesia che ne conclude la *princeps* (*Paesaggio* [VI]) può essere messa alla prova dallo specialista o rubare l'occhio allo studente, per la visione di prima mano dell'officina alleggerita dalle difficoltà di una lettura non sempre semplice dei segni d'autore.

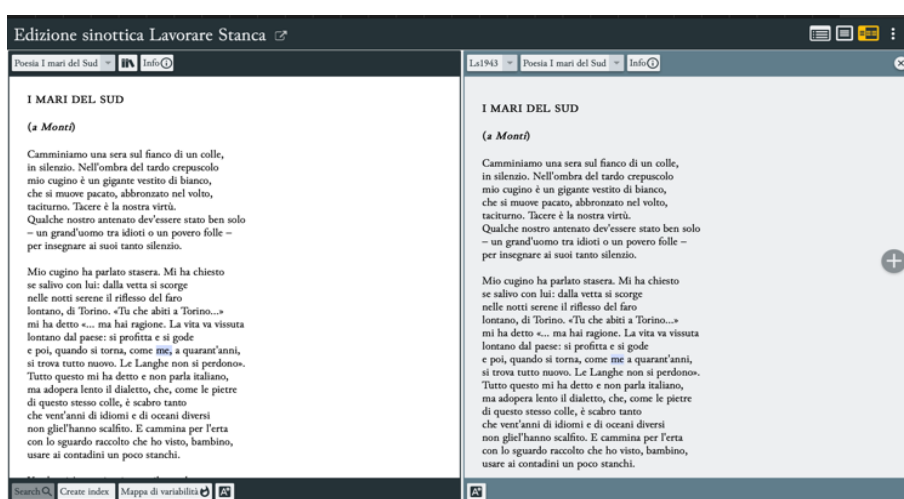


Figura 2. Edizione sinottica di *Lavorare Stanca* 1936-1943.

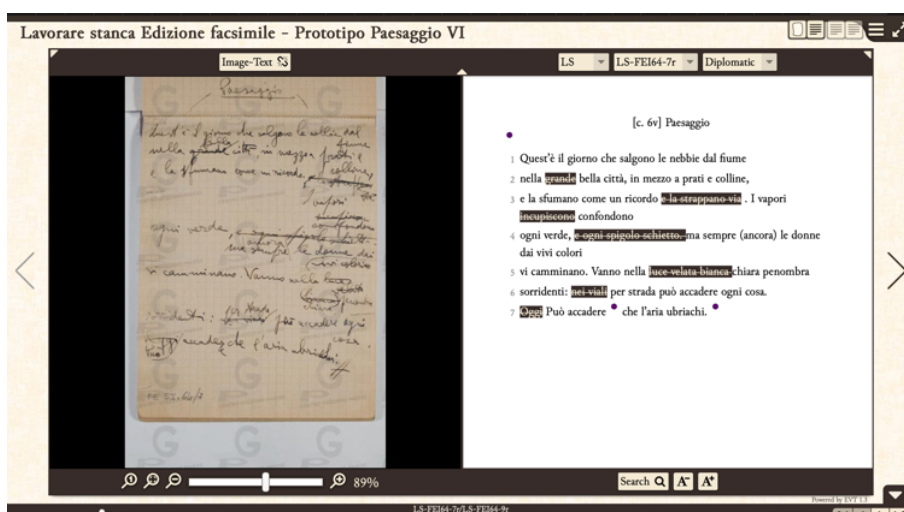


Figura 3. Prototipo di edizione facsimile (*Paesaggio* VI).

Se *Lavorare stanca* diventa il testo esemplare su cui mettere alla prova l'edizione sinottica e quella facsimile – in quanto testo fondamentale della poesia contemporanea, libero dai diritti d'autore –, *Il nome della rosa* rappresenta invece il modello di un discorso alternativo su testi ancora protetti dal *copyright* (condizione tipica di gran parte della letteratura del Novecento). Su *Il nome della rosa* sono state quindi allestite due diverse edizioni: l'edizione apparato che presenta una selezione di varianti tra il testo del 1980 e quello del 2012 e l'edizione intertestuale (Figura 4), basata su [1] e sulle

integrazioni dell'editore critico, che rilegge i tre *incipit* de *Il nome della rosa* utilizzando *tag* quali <quote> e <seg> per visualizzare tramite EVT2 tutto l'apparato intertestuale (dalla citazione esplicita di *Gv* 1,1 al passo parallelo dei *Peanuts*). Accanto a ciò si pone la concordanza de *La luna e i falò* (Figura 5) che si basa sul sistema di lemmatizzazione LiotroConcord [4], il quale consente: a) la lemmatizzazione integrale del testo base (nel nostro caso l'edizione curata da Grasso [5]); b) la quantificazione di tutte le occorrenze di parola, forme e lemmi; c) la produzione di liste lessicali strutturate grammaticalmente, formalmente e statisticamente; d) la creazione di diversi tipi di liste di frequenza (decescenti, per categorie grammaticali, alfabetiche, ecc.).

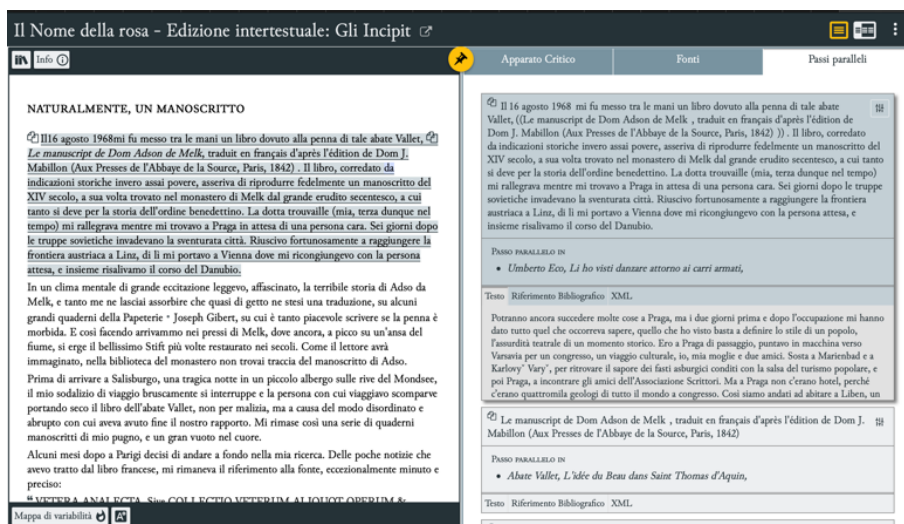


Figura 4. Prototipo di edizione intertestuale *Il nome della rosa*.



Figura 5. Concordanza de *La luna e i falò*.

### 3. FUORICLASSE

Nella sezione «FuoriClasse» si trovano poi tutte le risorse multimediali a corredo delle opere trattate scientificamente nell'Edizione-archivio: percorsi didattici, tematici e itinerari di senso; podcast; graphic novel; apparato iconografico e audio-video; materiali concordanziali e data visualization. L'orientamento didattico basato sulla psicologia della Gestalt e sull'approccio umanistico-affettivo è descritto in [2] e mette al centro il lettore e i suoi bisogni educativi con l'obiettivo di integrare scientificità (dalle varianti d'autore all'impianto lessicografico dell'opera) e *storytelling* con dei prototipi di *graphic novel* sulla vita di Pavese (Figura 6) e di podcast sulla vita dello stesso (Figura 7).





Figura 6. Graphic novel su Pavese.

CATABASI. I GIORNI DEL CONFINO - PODCAST

EPISODIO 1



EPISODIO 2



Figura 7. Podcast sulla vita di Pavese.

#### 4. SVILUPPI FUTURI

LINEA si configura come progetto pilota sulla Letteratura Italiana del Novecento, rivolto a studiosi e studenti, che mette a frutto l'esperienza maturata da parte dell'équipe del CINUM in ALPION e in PirandelloNazionale. Sarà sicuramente da affrontare la questione che riguarda la possibilità di aggregare esperienze simili sia verso un ampliamento del corpus di testi e autori (Calvino, Ungaretti, Pasolini, eccetera), sia attraverso l'integrazione dei portali monografici sugli autori già trattati, in linea con il Piano Nazionale di Digitalizzazione del Patrimonio Culturale e grazie alle opportunità del PNRR. Ad esempio, è attualmente in progettazione un portale *Digital Pavese* che presenti tutti i materiali manoscritti di Pavese, le edizioni sinottiche più importanti, la lemmatizzazione integrale delle opere maggiori, proponendo studi che si servano del *Natural Language Processing* per indagare il rapporto esistente tra le due edizioni di *Lavorare stanca*. È anche in elaborazione *Digital Eco*, uno spazio digitale di approfondimento sulle opere di Umberto Eco sia da un punto di vista filologico (dialogando con esperienze quali *Le varianti della rosa* [3]) sia lessicografico (con lo sviluppo dei vocabolari delle opere di Eco e di studi quantitativi e qualitativi sul lessico dell'autore alessandrino).

#### BIBLIOGRAFIA

- [1] Barbarino, Liborio Pietro. *Il primo Lavorare stanca di Pavese (1936)*. Edizione critica, Avellino: Edizioni Sinestesie, 2020.
- [2] Contini, Gianfranco. "Come lavorava l'Ariosto". In *Esercizi di lettura*, Torino: Einaudi, 1982.
- [3] D'Agata, Christian. "L'edizione scientifica digitale estesa de «Il nome della rosa»: modellizzazione, workflow e il paradigma IDEA", *Umanistica digitale*, n.14, 2022.
- [4] Di Silvestro, Antonio, Christian D'Agata, Giuseppe Palazzolo, Pietro Sichera, "Conservazione e fruizione di banche dati letterarie: l'archivio della poesia italiana dell'Otto/Novecento di Giuseppe Savoca". In *AIUCD 2022 - Proceedings*. A cura di: F. Ciraci, G. Miglietta, C. Gatto, 2022.
- [5] Eco, Umberto. *Il nome della rosa*. Introduzione e note a cura di Costantino Marmo. Milano: Bompiani, 1990.
- [6] Genette, Gérard. *Palinsesti. La letteratura al secondo grado*. Torino: Einaudi 1997.
- [7] Giuffrida, Milena, Christian D'Agata, Laura Giurdanella, Pietro Sichera. "Pirandello Nazionale: per un nuovo modello di edizione digitale, collaborativa e integrata". In F: Boschetti, A. M. Del Grosso, E. Salvatori (edd.), *AIUCD 2021- DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale. Raccolta degli abstract estesi della 10° conferenza nazionale*, 2021.
- [8] Kristeva, Julia. *Semiotiké. Ricerche per una semanalisi*. Milano: Feltrinelli, 1978
- [9] Mancinelli, Tiziana, e Elena Pierazzo. *Che cosa è un'edizione scientifica digitale*. Roma: Carocci, 2020.
- [10] Pavese, Cesare. *La luna e i falò*. Edizione critica a cura di Miryam Grasso. Avellino: Edizioni Sinestesie, 2020.
- [11] Rosselli Del Turco, Roberto, Chiara Di Pietro, Chiara Martignano. "Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo", *Umanistica Digitale*, n. 7, 2019.
- [12] Sahle, Patrick. "What is a Scholarly Digital Edition?". In *Digital Scholarly Editing. Theories and Practices*. Edited by M. J. Driscoll and E. Pierazzo, 19-40. Cambridge, UK: Open Public Publishers, 2016.
- [13] Savoca, Giuseppe. *Lessicografia letteraria e metodo concordanziale*. Firenze: Olschki, 2000.
- [14] Stoppelli, Pasquale, e Eugenio Picchi. *LIZ 4.0. Letteratura italiana Zanichelli*. Bologna: Zanichelli, 2001.
- [15] TEI Consortium, eds. *Guidelines for Electronic Text Encoding and Interchange*. [25th October 2022]. <http://www.tei-c.org/P5/>



# ‘Voce dei Margini’: Le annotazioni di Giorgio Bassani

Angela Siciliano<sup>1</sup>, Vanessa Bianconi<sup>2</sup>, Angelo Mario Del Grosso<sup>3</sup>

<sup>1</sup> Université Sorbonne Nouvelle-Paris 3, Francia – angela.siciliano93@gmail.com

<sup>2</sup> Università di Pisa, Italia – v.bianconi@studenti.unipi.it

<sup>3</sup> Istituto di Linguistica Computazionale ‘A. Zampolli’ (CNR-ILC), Italia – angelo.delgrosso@ilc.cnr.it

## ABSTRACT

Il contributo presenta gli avanzamenti del progetto di edizione scientifica finalizzato alla rappresentazione digitale e alla fruizione web del *corpus* di postillate annotate dallo scrittore Giorgio Bassani sui libri della sua biblioteca personale. Il progetto, che abbiamo denominato “La voce dei Margini di Giorgio Bassani”, si è articolato in due fasi principali: la prima, di modellazione e codifica dell’edizione digitale, ha fatto uso delle norme proposte dalle linee guida XML-TEI; la seconda fase è stata dedicata allo sviluppo di un *applet* – denominata “Voce dei Margini” – finalizzata alla presentazione e all’interrogazione del *corpus* codificato. Ad oggi il prototipo di edizione, modellato sul caso esemplare del volume *La scuola dell’uomo* di Guido Calogero, ha raggiunto uno stadio di maturità e di generalità tale da poter accogliere anche altri volumi postillati da Bassani.

## PAROLE CHIAVE

Postille d’autore; Digital Scholarly Edition; Filologia Computazionale; eXist-db; Giorgio Bassani

## 1. INTRODUZIONE

Giorgio Bassani (1916-2000) è oggi riconosciuto come un autore di primo piano, per la ricchezza e la versatilità del suo profilo: narratore, poeta, saggista, editor, sceneggiatore, professore, attivista per l’ambiente, Bassani ha segnato profondamente il Novecento letterario italiano con il suo stile, la sua figura, le sue scelte poetiche ed editoriali. Da qui la necessità di ripercorrerne nel dettaglio la biografia intellettuale [1,2], che riflette – oltre il dato strettamente personale – anche le tendenze, i gusti e gli orientamenti culturali di un’epoca. Uno strumento privilegiato per condurre questa indagine è senza dubbio la biblioteca personale dello scrittore, oggi conservata presso la Fondazione Giorgio Bassani di Ferrara: un ricco patrimonio (circa 3400 volumi) impreziosito dalla fitte annotazioni d’autore. Ed è proprio a partire dall’esame delle postille – scritture marginali, parassitarie e apparentemente minori, in realtà dal grande potenziale critico – che è possibile indagare, con percorsi inediti e sorprendenti, la genesi del pensiero, della poetica e delle opere di Bassani. Un potenziale critico, quello delle postille d’autore, che può essere valorizzato nell’ambiente digitale, che consente una più ordinata rappresentazione e fruizione della messe di dati [3,4] e potenzialmente un accesso aperto e ubiquo ad essa [5], come dimostrano alcuni progetti nazionali e internazionali sul tema: *AMARGINE* [6], *Beckett Digital Library*<sup>1</sup>, *Biblioteca di Manzoni*<sup>2</sup>, *In Whitman’s Hand. Annotations and Marginalia*<sup>3</sup>, *Melville Marginalia Online*<sup>4</sup>, *Mill Marginalia Online*<sup>5</sup>. Si è scelto così di allestire l’edizione scientifica digitale delle postille di Bassani, un progetto che si snoda lungo molteplici linee di ricerca scientifica e di sviluppo tecnologico [7]. La prima fase è stata dedicata alla modellazione dell’edizione digitale [8] attingendo, dal vocabolario XML-TEI<sup>6</sup> [9,10], gli elementi più appropriati per plasmare la struttura della risorsa, rappresentando al meglio il complesso ecosistema testuale, paratestuale ed extratestuale [11] costituito dall’opera postillatoria di Bassani<sup>7</sup>. La seconda fase ha riguardato quindi lo sviluppo di una applicazione web in ambiente XML-nativo eXist-db<sup>8</sup> orientata alla presentazione, fruizione e analisi del *corpus*. Nel presente contributo approfondiremo alcune funzionalità implementate durante le varie fasi del progetto di edizione digitale, descrivendone i principi di base sia per quanto riguarda i dettagli tecnologici sia per quelli di approfondimento scientifico-filologico.

Per illustrare il lavoro svolto faremo riferimento, come caso di studio, alla postillatura di una singola pagina (p. 100) del volume *La scuola dell’uomo* di Guido Calogero [13]: si è scelto infatti di partire da questo volume, esemplare per ricchezza

<sup>1</sup> <https://www.beckettarchive.org>

<sup>2</sup> <http://www.alessandromanzoni.org>

<sup>3</sup> <https://whitmanarchive.org/manuscripts/marginalia/index.html>

<sup>4</sup> <https://melvillesmarginalia.org/>

<sup>5</sup> <https://millmarginalia.org/>

<sup>6</sup> <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

<sup>7</sup> Altrove sono state analizzate nel dettaglio le scelte di codifica effettuate per la rappresentazione digitale delle postille. Nel presente contributo ci limiteremo solo a sintetizzare i principi guida e gli elementi più importanti [12].

<sup>8</sup> <http://exist-db.org/exist/apps/homepage/index.html>

e completezza della fenomenologia postillatoria (circa 200 postille), per mettere a punto un modello di edizione estendibile senza alcun adattamento sostanziale all'intera biblioteca dello scrittore ferrarese (circa 3400 volumi, di cui 250 postillati).

## 2. IL PROGETTO DIGITALE “LA VOCE DEI MARGINI DI GIORGIO BASSANI”

Il modello di rappresentazione digitale del testo, definito nella prima fase del progetto, prevede un'articolazione multidimensionale dell'oggetto postilla. Infatti sono stati considerati diversi piani: il piano del documento facsimile, il piano della trascrizione diplomatica, il piano dell'edizione critica e, infine, il piano del testo a stampa *target* delle annotazioni d'autore. Accanto alle dimensioni testuali e documentarie, il modello predisposto per l'edizione scientifica digitale delle postille di Bassani propone una inversione della gerarchia testo-paratesto tra il contenuto a stampa e il corrispondente intervento autoriale: il testo principale diviene la postilla, mentre il testo a stampa, annotato dallo studioso, assume il ruolo di paratesto. Questo principio di inversione si è tradotto concretamente in opportune scelte catalografiche, descrittive, classificatorie e analitiche espresse per mezzo degli elementi e degli attributi definiti nel vocabolario XML-TEI. In particolare, l'uso “originale” del tagset definito nel modulo 10 delle linee guida<sup>9</sup> relativo alla descrizione della fonte primaria. A ciò, occorre aggiungere la registrazione dei collegamenti alle regioni di interesse presenti sulla riproduzione facsimilare (come definito dal modulo 11 delle linee guida TEI<sup>10</sup>) nonché i rimandi alle classi della tassonomia all'uopo definita (seguendo il modello descritto nel modulo 2 delle linee guida<sup>11</sup>) utilizzato per la classificazione delle postille secondo tre diverse categorie di base: verbali, non verbali e miste (cioè verbali e non verbali insieme).

Come mostra il caso di pagina 100 della *Scuola dell'uomo* (Fig. I-app), possiamo esemplificare le scelte di codifica (Fig. 1) partendo dall'entrata catalografico-archivistica della postilla con gli elementi del tagset <msContents>, in particolare <msItem>, <locus>, <textLang>, <note>, passando poi per la codifica dell'edizione documentaria mediante il tagset <sourceDoc>, in particolare gli elementi <line>, <zone>, <metamark>, per giungere alla codifica dell'edizione critica facendo uso degli elementi strutturali <text>, <div> e <p>. Infine è stato trascritto anche il testo dell'edizione a stampa annotato dall'autore, in particolare con gli elementi strutturali <text>, <div> e <ab><sup>12</sup>. Le postille sono tutte ordinate numericamente e condividono uno spazio reticolare di relazioni tra il documento, il testo e i riferimenti incrociati.

<pre>&lt;msItem n="103" xml:id="p103" class="#postilla_verbale_legata_a_segno_non_verbale #segno_non_verbale" facs="#pag100.p2"&gt; ...&lt;msItem n="103.1" class="#postilla_personale #postilla_polemica #postilla_di_filosofia #postilla_di_pedagogia"&gt; ...&lt;locus target="#t100.2" facs="#pag100.p2det"&gt; p. 100 &lt;/locus&gt; ...&lt;textLang&gt; italiano &lt;/textLang&gt; ...&lt;note resp="#AS"&gt; postilla verbale a lapis, sul margine sinistro &lt;/note&gt; ...&lt;/msItem&gt; ...&lt;msItem n="103.2" class="#sottolineatura_interlineare_snv"&gt; ...&lt;locus facs="#pag100.p2det1"&gt; p. 100 &lt;/locus&gt; ...&lt;note resp="#AS"&gt; sottolineatura interlineare a lapis &lt;/note&gt; ...&lt;/msItem&gt; &lt;/msItem&gt;</pre>	<pre>&lt;zone xml:id="pag100.p2det" ulx="246" uly="1099" lrx="503" lry="1488" hand="#m1" corresp="#pag100.p2t"&gt; ...&lt;line&gt; &lt;metamark function="division_of_notes" rend="asterisk" place="between_the_notes"/&gt; Ma allora è fare &lt;/line&gt; ...&lt;line&gt; storia l'atto di imparziale ob &lt;/line&gt; ...&lt;line&gt; biettività, e &lt;/line&gt; ...&lt;line&gt; solo questo; o &lt;/line&gt; ...&lt;line&gt; anche il succes &lt;/line&gt; ...&lt;line&gt; sivo etico-peda &lt;/line&gt; ...&lt;line&gt; gogico di valu &lt;/line&gt; ...&lt;line&gt; tazione etica. &lt;/line&gt; ...&lt;/zone&gt;</pre>
<pre>&lt;div n="103" corresp="#p103" xml:id="t100.2" type="postilla" subtype="marginalia"&gt; ...&lt;p&gt; Ma allora è fare storia l'atto di imparziale obbiettività, e solo questo; o anche il successivo etico-pedagogico di valutazione etica. &lt;/p&gt; ...&lt;/div&gt;</pre>	

Figura 1. Scelte di codifica per la registrazione dei metadati, per la rappresentazione documentale e per il testo critico (frammento XML-TEI postilla n. 103 pagina 100 del volume la *Scuola dell'uomo*).

## 3. PRESENTAZIONE E FRUIZIONE DEL CORPUS: L'APPLET “VOCE DEI MARGINI”

Dopo aver definito il modello del testo e codificato il *corpus* delle postille alla *Scuola dell'uomo*, il lavoro di edizione scientifica digitale si è orientato alla progettazione e alla realizzazione di un'applicazione ad accesso web mediata da *Graphical User Interface* (GUI). L'ambiente web è stato disegnato per essere funzionale alla presentazione dei dati codificati in una modalità che fosse il più possibile congeniale – per qualità, quantità e fruibilità delle informazioni sulle

<sup>9</sup> <https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>

<sup>10</sup> <https://tei-c.org/release/doc/tei-p5-doc/en/html/PH.html>

<sup>11</sup> <https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html>

<sup>12</sup> Seppur monotestimoniale, il modello di codifica prevede sia la rappresentazione della trascrizione documentaria (blocco sourceDoc) sia la rappresentazione dell'edizione critica (blocco text), rispondendo in parte alle obiezioni sollevate in [14] sulla predominanza di edizioni digitali puramente documentarie, dove l'editore si esime dall'esprimere un giudizio critico per la mediazione del testo finale.

postille – al filologo interessato allo studio del pensiero e dell’opera di Bassani. Lo sviluppo dell’*applet* si compone di una prima fase durante la quale sono stati raccolti i requisiti, stabilendone gli scopi e le principali funzionalità. Abbiamo proceduto quindi all’implementazione della sua prima realizzazione in forma di prototipo. La fruizione dell’edizione, sviluppata mediante il modulo *HTML templating library* fornito dalla piattaforma eXist-db, è un’*applet* web finalizzata alla pubblicazione *open-access* e all’interrogazione avanzata di tutto il *corpus* della biblioteca di Bassani (Fig. 2)<sup>13</sup>. Di particolare interesse per la sua originalità è l’uso dell’ambiente XML-nativo, adottato per realizzare l’intera fruizione dell’edizione (prendendo a modello il progetto di edizione digitale delle lettere di Van Gogh<sup>14</sup>, basato sull’applicativo TEIPublisher<sup>15</sup>). In questo modo è infatti possibile disporre di tutto l’occorrente tecnologico volto all’interrogazione, all’elaborazione e alla presentazione dei dati codificati. La tecnologia di riferimento per lo sviluppo dell’*applet* è il linguaggio d’interrogazione XML XQuery<sup>16</sup>.



Figura 2. Presentazione e selezione dei volumi postillati da Giorgio Bassani.

L’applicazione di filologia computazionale<sup>17</sup>, che abbiamo chiamato “Voce dei Margini”<sup>18</sup>, presenta integralmente il *corpus*, mostrando le postille d’autore e il corrispondente testo a stampa annotato. Dall’*home page* del progetto è possibile scegliere il volume da analizzare (Fig. 2). Il modello elaborato sulla *Scuola dell’uomo* è estendibile, per struttura e funzionalità, a tutti i volumi postillati da Bassani: abbiamo in programma, come prossimo avanzamento del lavoro, di aggiungere all’attuale *corpus* i volumi su cui lo scrittore ha elaborato la sua coscienza politica (tra il 1937 e il 1945 militò, prima attivamente poi clandestinamente, nell’antifascismo, scontando anche il carcere e l’esilio da Ferrara). Una volta individuato il volume è possibile poi selezionare, pagina per pagina, la riproduzione facsimilare dove sono presenti le note d’autore (Fig. 3). La schermata offre una rappresentazione e una descrizione esaustiva della pagina postillata: in essa ritroviamo la scansione del volume a stampa (presentato per mezzo del componente di visualizzazione d’immagini javascript OpenSeadragon<sup>19</sup>), la trascrizione delle postille verbali, l’indicazione degli interventi d’autore e la descrizione delle postille non verbali e dei collegamenti tra le postille e il testo annotato, nonché i dettagli descrittivi e catalografici del volume.

Il contenuto di ogni singola postilla è stato analizzato automaticamente dal punto di vista linguistico mediante il modulo di trattamento automatico della lingua sviluppato dall’Università di Stanford<sup>20</sup> e integrato nella piattaforma eXist-db<sup>21</sup>. Il modulo di analisi automatica della lingua ci permette di ricavare lemma, categoria grammaticale e presenza di entità nominate di tutte le parole che compongono la postilla<sup>22</sup>.

<sup>13</sup> Il prototipo dell’*applet* in sviluppo è raggiungibile al seguente link: <https://48002.omega.ilc.cnr.it/exist/apps/postille/index.html>

<sup>14</sup> <https://teipublisher.com/exist/apps/vangogh/index.html>

<sup>15</sup> <https://teipublisher.com/index.html>

<sup>16</sup> <https://www.w3.org/TR/xquery-31/>

<sup>17</sup> Con il termine filologia computazionale ci riferiamo allo sviluppo e/o all’utilizzo di componenti software volti ad automatizzare sia le procedure per l’edizione testuale, sia le procedure per l’analisi del testo e della lingua nonché l’estrazione di dati testuali, sia le procedure di indicizzazione e di ricerca dei contenuti [15].

<sup>18</sup> L’applicazione sviluppata sarà disponibile con licenze aperte su piattaforme di condivisione software quali Github.

<sup>19</sup> <https://openseadragon.github.io/>

<sup>20</sup> <https://stanfordnlp.github.io/CoreNLP/>

<sup>21</sup> <https://github.com/eXist-db/exist-stanford-nlp>

<sup>22</sup> Il modello utilizzato per l’elaborazione e l’annotazione linguistica nel contesto del prototipo “Voce dei Margini” è stato sviluppato in seno al progetto TINT [16] che estende e completa per la lingua italiana gli strumenti sviluppati dalla libreria CoreNLP dall’Università di Stanford.

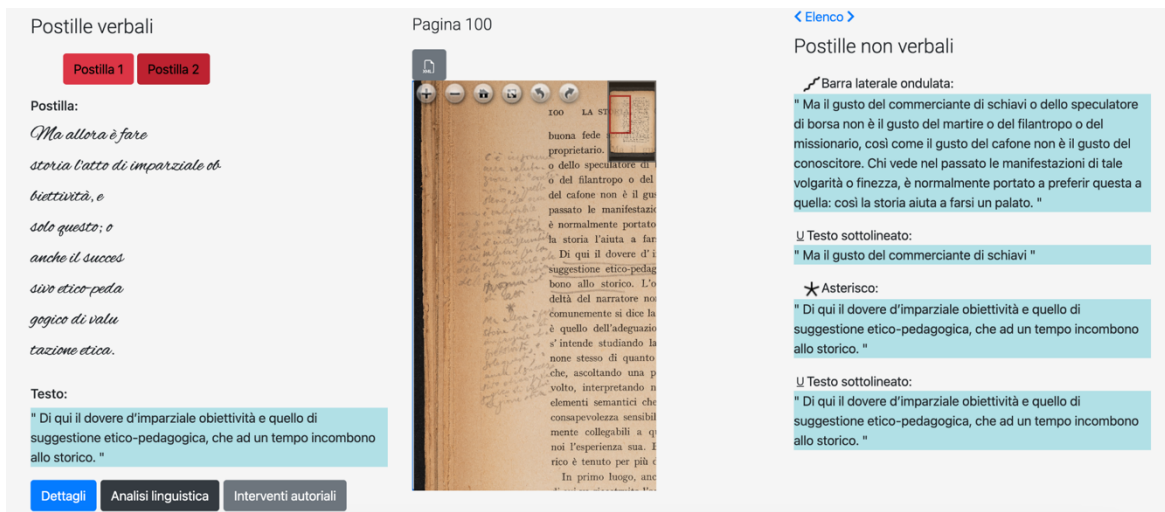


Figura 3. Dettaglio postille verbali, non verbali e miste di pagina 100 del volume *la Scuola dell'uomo* di G. Calogero.

Tra le potenzialità più interessanti dell'applicazione "Voce dei Margini", si segnala la possibilità di fornire allo studioso un agile, ma potente strumento di interrogazione del *corpus* lungo diverse modalità e direttrici di ricerca. Ad oggi sono possibili quattro diverse tipologie di interrogazione: 1) ricerca wildcard, 2) ricerca fuzzy, 3) ricerca di prossimità, 4) ricerca entità nominate. L'esempio di Figura 4 mostra l'interrogazione del *corpus* delle postille mediante una ricerca di prossimità tra due parole che co-occorrono in una finestra di contesto definita dall'utente (nel caso specifico, il contesto definito da due parole inframezzate ai termini di ricerca: "storia" e "imparziale").

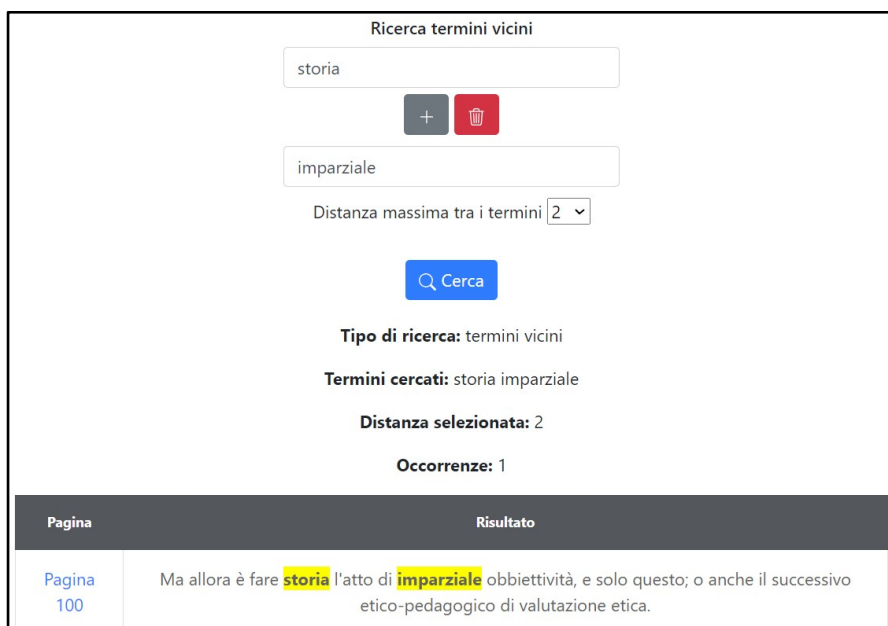


Figura 4. Interrogazione del *corpus* per mezzo dei termini "storia" e "imparziale" in una finestra di contesto di due parole.

La funzione di ricerca è implementata utilizzando il sistema di indicizzazione *full-text* disponibile tra i moduli di eXist-db, che fa uso del sofisticato motore di ricerca noto con il nome *Lucene*<sup>23</sup>. Grazie a questa tecnologia, particolarmente efficace e funzionale, è possibile interrogare il testo compilando "query" molto articolate seguendo una sintassi specifica definita dal linguaggio d'interrogazione. Un'ulteriore funzionalità d'indagine implementata all'interno dell'applet riporta una lista di termini notevoli predefinita, attraverso cui svolgere un percorso di approfondimento guidato nel *corpus* delle postille oggetto di studio. I termini selezionati sono quelli che compongono il lessico del giovane antifascista Bassani (Fig. 5). Per esempio, cliccando il pulsante con la parola "storia", l'applicazione restituisce tutte le pagine dove è presente una postilla che annovera tra le sue parole anche quella selezionata.

<sup>23</sup> <https://lucene.apache.org/>



Cruce		
Educazione		
Etica		
Liberalismo		
Libertà		
Morale		
Politica		
Spirito		
Storia		
Storicismo		
Parola selezionata: storia		
Occorrenze: 6		
Pagina	Risultato	Testo a stampa
Pagina 96	ma la <b>storia</b> è sempre prodotto, risultato, della volontà umana? La <b>storia</b> non travolge spesso la volontà? L'atto dell'uomo è sempre documento di una sufficiente consapevolezza? la <b>storia</b> non è qualche volta il segno di una irrazionalità irrefrenabile, il regno del dio cattivo?	ultime azioni dovettero pure esser sentite come preferibili ad ogni altra dalla presente coscienza di chi le compì, giacché, altrimenti, esse non sarebbero state compiute.
Pagina 96	la <b>storia</b> come educatrice Croce	Dal punto di vista di chi non ha più come agire, la storia non può presentarsi che nell'aspetto della pura comprensione, quale semplice ricostruzione sgiustificatrice dell'esperienza di chi agì
Pagina 100	Ma allora è fare <b>storia</b> l'atto di imparziale obiettività, e solo questo; o anche il successivo etico-pedagogico di valutazione etica.	Di qui il dovere d'imparziale obiettività e quello di suggestione etico-pedagogica, che ad un tempo incombono allo storico.

Figura 5. Indagare il corpus guidato da una selezione pre-costituita di termini notevoli antifascisti.

#### 4. CONCLUSIONI E PROSPETTIVE FUTURE

L'edizione scientifica digitale delle postille di Giorgio Bassani si è sviluppata in un'applicazione web denominata "Voce dei Margini". Il modello di testo-postilla, inizialmente elaborato durante le ricerche di dottorato di uno degli autori, è stato concretamente verificato su un primo volume (*La scuola dell'uomo* di Guido Calogero) per testarne l'efficacia. Recentemente l'edizione si è dotata anche di un prototipo per la fruizione e l'interrogazione della collezione di postillati codificati in XML-TEI. Il prototipo è stato sviluppato a partire dalle tecnologie disponibili all'interno della piattaforma XML-nativa eXist-db.

È nostra intenzione seguire i principi FAIR<sup>24</sup> per il rilascio del corpus, affidandoci ad autorevoli infrastrutture di ricerca di ambito umanistico quali CLARIN<sup>25</sup> e DARIAH<sup>26</sup> (le due infrastrutture sono ora parte del consorzio H2IOSC<sup>27</sup> per lo sviluppo di una piattaforma comune in seno alle iniziative dell'open science cloud europeo<sup>28</sup>). Per quel che concerne le licenze dei dati e per la tutela della proprietà d'autore, è in corso una trattativa con gli eredi di Bassani per la pubblicazione dei dati in una forma il meno possibile restrittiva. In conclusione, il contributo ha illustrato le innovazioni più recenti del progetto mettendo in evidenza le potenzialità critiche dello strumento digitale in sviluppo. I prossimi avanzamenti che abbiamo deciso di sviluppare riguardano in prima istanza la disponibilità di nuovi dati codificati, in particolare il testo delle postille di alcuni volumi che Bassani ha annotato per la formazione del suo pensiero politico (*La libertà nello stato moderno* di Harold J. Laski, 1931; *Teoria e storiografia* di Benedetto Croce, 1927; *Socialisti tedeschi* di Giorgio Pasquali, 1919; *Storia d'Europa* di H. Fisher, 1938). Dal punto di vista applicativo, considereremo nuove funzionalità da aggiungere all'applet, in particolare un collegamento ottimizzato tra testo e immagine e un motore di ricerca che possa includere tutti i metadati dei volumi annotati e delle postille. In più, stiamo pensando a funzionalità di analisi automatica del testo per raggruppare automaticamente postille simili oppure postille con specifiche caratteristiche linguistico-testuali.

#### BIBLIOGRAFIA

- [1] Bausi F., "Bassani prima di Bassani. Preistoria di un intellettuale e di un narratore (1935-1942)", in *Il tempo dello spirito. Saggi per il centenario della nascita di Giorgio Bassani*, a cura di A. Perli, «Sinestesia», 2016, pp. 21-43.
- [2] Scarpa D., "Sentieri interrotti e sentieri battuti. Giorgio Bassani 1935-1943", in *Bassani nel suo secolo*, a cura di S. Amrani e M. P. De Paulis-Dalembert, Ravenna, Giorgio Pozzi Editore, 2017, pp. 33-74.
- [3] Estill, L. 2016. "Encoding the Edge: Manuscript Marginalia and the TEI." *Digital Literary Studies* 1 (1).
- [4] Boot, P. 2009. *Mesotext: Digitised Emblems, Modelled Annotations and Humanities Scholarship*. Amsterdam: Pallas Publications.
- [5] Cummings, J. 2019. "Opening the Book: Data Models and Distractions in Digital Scholarly Editing." *International Journal of Digital Humanities* 1 (2): 179-93. <https://doi.org/10.1007/s42803-019-00016-6>
- [6] Milone F., Postille "amargine": problemi e metodi per l'edizione digitale dei postillati di autori del secondo Novecento, «Linguistica e Letteratura», 1-2 (2022), pp. 195-209.

<sup>24</sup> <https://www.go-fair.org/fair-principles/>

<sup>25</sup> <https://www.clarin-it.it/it>

<sup>26</sup> <http://it.dariah.eu/>

<sup>27</sup> <https://www.h2iosc.cnr.it/>

<sup>28</sup> <https://eosc-portal.eu/>

- [7] Boschetti, F., and A. M. Del Grosso. 2020. "L'annotazione di testi storico-letterari al tempo dei social media." *Italica Wratislaviensia* 11 (1): 65–99.
- [8] Buzzoni, M. 2016. "A Protocol for Scholarly Digital Editions? The Italian Point of View." In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 59–82. Open Book Publishers.
- [9] Pierazzo, E. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Digital Research in the Arts and Humanities. Farnham Surrey: Ashgate.
- [10] Burnard, L. 2014. What Is the Text Encoding Initiative? *Encyclopédie Numérique* 3. Marseille: OpenEdition Press.
- [11] Mancinelli, T., and E. Pierazzo. 2020. *Che Cos'è Un'edizione Scientifica Digitale*. Roma: Carocci Editore.
- [12] Siciliano, A., and A. M. Del Grosso. 2022. "From Print to Digital: An Encoding Model for the Scholarly Edition of Giorgio Bassani's Notes." *Umanistica Digitale* 13.
- [13] Calogero G., *La scuola dell'uomo*, Firenze, Sansoni, 1939.
- [14] Nava, B. 2022. "Are We All Bédierian? Perspectives for Digital Genetic Editions." *Umanistica Digitale* 6 (14): 19–40. <https://doi.org/10.6092/issn.2532-8816/14949>
- [15] Bozzi, A., M.S. Corradini, and G. Ferrari. 2019. *Percorsi Di Linguistica e Di Filologia Computazionali*. Edizioni ETS.
- [16] Palmero Aprosio, A., and G. Moretti. 2016. "Italy Goes to Stanford: A Collection of CoreNLP Modules for Italian." *ArXiv E-Prints*, September

buona fede sconfitta): e in tal senso giustifica il suo proprietario. Ma il gusto del commerciante di schiavi o dello speculatore di borsa non è il gusto del martire o del filantropo o del missionario, così come il gusto del cafone non è il gusto del conoscitore. Chi vede nel passato le manifestazioni di tale volgarità o finezza, è normalmente portato a preferir questa a quella: così la storia l'aiuta a farsi un palato.

Di qui il dovere d'imparziale obiettività e quello di suggestione etico-pedagogica, che ad un tempo incombono allo storico. L'obiettività, l'imparzialità, la fedeltà del narratore non sono altra cosa da quella che comunemente si dice la verità storica: il suo criterio è quello dell'adeguazione ermeneutica, la quale (come s'intende studiando la natura del linguaggio) è il canone stesso di quanto sempre operiamo tutte le volte che, ascoltando una persona parlare, studiando il suo volto, interpretando nel loro complesso tutti gli altri elementi semantici che nell'intero ambito della nostra consapevolezza sensibile o mnemonica ci appaiono utilmente collegabili a quelli, cerchiamo di realizzare in noi l'esperienza sua. E a tale dovere di verità lo storico è tenuto per più di un motivo.

In primo luogo, anche nel caso in cui la personalità di cui va ricostruita l'azione non sia più vivente, e quindi, ormai incapace di dolore e di gioia, non costituisca più il termine di una diretta obbligazione morale da parte dello storico, quest'ultimo deve pur sentire di fronte ad essa quello stesso dovere di umile adeguazione e in-

*C'è un'idea  
aria voluta  
giorno di conte  
nuto; quello  
stesso da non  
ma è volubile  
in g  
ai orologio  
e di  
è un'idea  
l'idea  
salvo  
della  
del*

*Ma allora è forse  
storia l'idea  
imparziale  
fedeltà  
solo questo  
anche il successo  
suo obiettivo  
magico di  
azione*

Figura 1 - Appendice. Postillatura di una singola pagina (p. 100) del volume *La scuola dell'uomo* di Guido Calogero.



# Scholarly edition of born-digital literary sources: a case study

Elena Spadini<sup>1</sup>

<sup>1</sup> University of Basel, Switzerland – elena.spadini@unibas.ch

## ABSTRACT

This paper reports on a case study of genetic edition of born-digital sources and reflects on the changes that this kind of materials bring to the philological work.

## KEYWORDS

Scholarly editing; born-digital sources; archives; digital philology

## 1. INTRODUCTION

Born-digital sources are nowadays ubiquitous in the cultural heritage field and beyond: textual scholarship, as well as any other approach to scholarly research on contemporary materials, will soon be impossible without engaging with born-digital sources. As the authors of the 2010 report from the Council and Information Resources on Digital Forensics and Born-Digital Content in Cultural Heritage Collections wrote, “[m]ost records today are born digital, and libraries and other collecting institutions increasingly receive computer storage media as part of their acquisition of “papers” from writers, scholars, scientists, musicians, and public figures” [5].

Born-digital sources, like any other source, should be archived and preserved, described, edited and analysed, in the context of archival and scholarly work. In Italy, the challenges posed by the archival of born-digital literary sources have been addressed in a series of contributions from the PAD (Pavia Archivi Digitali) project [1, 9, 10]. The archiving, description and analysis of born-digital literary sources are also the subject of the recently published book *Digitale d'autore. Macchine, archivi, letteratura* by Emmanuela Carbé [2], which provides a detailed account of the history and development of PAD and a thorough analysis of one of the PAD collections, that of the author Francesco Pecoraro. Based on ten years of research in the field and on an extremely rich bibliography, the book touches on the most important aspects of the work that scholars and archivists have to face when dealing with born-digital (literary) sources.

In this paper, I would like to focus on just one facet of this complex scenario, namely the editing of born-digital sources. Scholarly editing is, of course, linked to archival description and preservation on the one hand and analysis on the other, but in what follows I will confine myself as much as possible to the editorial tasks.

Similar to what is happening in other disciplines, the philological workflows and methods may be reshaped when addressing born-digital sources. Kirschenbaum and Reside, among the key figures in the study of born-digital literary archives, claimed that “new textual forms require new work habits, new training, new tools, new practices, and new instincts” [6]. The born-digital source should be transformed into a scientific object of study.

## 2. A CASE STUDY

The case study presented in this paper is a prototype of the genetic edition of (a part of) *Tout passe* (Christan Bourgois éditeur, 2011), a collection of short stories by the Swiss author Bernard Comment, which won the French Prix Goncourt de la Nouvelle in 2011. The archives of Bernard Comment were acquired by the Swiss Literary Archives (SLA), part of the Swiss National Library, in 2013 and on subsequent occasions. The acquisition includes literary material, as well as documents relating to his work in the cultural sector and his correspondence (for example with Antonio Tabucchi, whom Comment translated extensively). The 139 archive boxes also contain digital documents on a hard disk, with files from 1998 to 2018. For the collection of short stories *Tout passe*, both paper (manuscripts, typescripts) and digital sources exist. I am in contact with the author through the archive and he is kept informed of any progress in the research.

The aim of the case study and of the prototype is to provide concrete examples for what can be widespread issues in scholarly editing of born-digital and mixed sources.

### 3. COLLECTION OF WITNESSES

The creation of a comprehensive collection of sources is the first step in any editorial workflow. In genetic editing and *filologia d'autore*, this phase corresponds to the creation of the genetic dossier.

The current situation in many archives does not facilitate this task, as Kirschenbaum and Reside's report (on the collection of the composer Jonathan Larson at the Library of Congress) shows: "Initially the Library had considered treating Larson's computer diskettes much like the other media in its audio collections –that is, catalog them according to the label on the object but without detailed listings of the files stored on them. This is not an unusual approach, and, given the strained budgets and workloads of most archives, along with the time and specialized technical expertise required for more detailed cataloging, it may be all that is possible in many cases" [6].

In the case of Bernard Comment at the SLA, the difficulties are similar. While the paper archive has been catalogued and metadata created for each document, the digital part has not yet been processed: the only information found in the catalogue concerns the entire hard-drive and provides a brief description of its contents, as well as a list of the main folders (<https://www.helveticaarchives.ch/detail.aspx?ID=1332712>). In the SLA reading room, the researcher is given a copy of the original hard disk on which all the documents have been converted to the PDF/A archiving format, retaining the original extension in the file name (example 'Un fils.docx.pdf'). The conversion to PDF/A is currently used in other Swiss archival institutions: in the Centre des littératures en Suisse romande (University of Lausanne), for example, the originals are converted and only the PDF/A version is archived. This archiving choice has consequences for the edition and the analysis of the sources, as we shall see.

In the absence of a detailed catalogue, and as file names cannot always be expected to be representative of content, the researcher must open each file individually to find potentially relevant material. Search functions should help, but difficulties may be caused by the operating system of the computer available in the reading room: depending on the operating system, GUI search functions and terminal commands vary significantly.

The external hard-drive where the digital sources are stored is unlikely to contain data to be recovered (with forensic methods), because it has been created at the moment of the acquisition and has never been used as a working device.

Once the witnesses have been identified, they should be arranged in chronological order. If "chronology is the backbone of the genetic edition" [10], the classification of digital sources might bring an advantage over paper ones, since metadata should make it possible to date and order versions with greater certainty. Indeed, the distributed materiality of digital objects is important here: it can happen, for example, that a system clock is incorrectly set [6]. Furthermore, it is not always easy, nor possible, to access the original metadata. In our case, we first have access only to the date of the last modification of the PDF/A converted files, which is the same for all the files on the hard-drive. After asking to the SLA archivists, though, they were able to provide more data (obtained with the DROID software, see <https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>), including the last modification date of the original files on the hard-drive.

### 4. ENCODING AND COLLATION

If the author has not used track-change functionality, and no digital forensic method has been applied, the sources are unlikely to contain traces of the writing process: as it has been widely noted, especially by early adapters of word processors, the text on the screen is always clean. Most of the TEI Guidelines chapter on the representation of primary sources (Chapter 11), which contains some of the most commonly used tags in the context of genetic editing, becomes irrelevant: there is no surface to map, no deletion or substitution, no ambiguous reading, and probably fewer abbreviations. The structure of the text is what remains to be encoded, along with information about formatting.

The next step in the editing process is to compare the states of the text, i.e. to collate them. Collation can be central to the edition of born-digital sources because the texts are already in digital form and do not need to be transcribed as a preliminary step: the non-availability of full transcriptions has severely limited the application of automatic collation to textual sources, but this problem disappears with born-digital sources. Moreover, the rewritings between versions identified by collation are the only evidence we have, since, as said, immediate rewritings are in many cases lost forever: in the words of Jean-Louis Lebrave, "[l]a génétique in forensique risque donc de devoir définitivement renoncer à être une poétique des processus pour se contenter d'être une poétique des transitions entre états", in which "le temps n'est pris en compte que sous forme discrétisée" [8].

In our case study, the structural encoding can help to compare the different witnesses, as collation programs perform better when aligning short texts. A macro-alignment (manual or computational) is an important preliminary step in all cases where the witnesses differ considerably. In the case of *Tout passe*, the witnesses are not many, but they may contain very different texts, that should be prepared for an effective collation. However, this problem is not limited to born-digital sources: a

possible influence of word processors on the extent of rewriting practices could only be established from the data on many authors and dossiers.

## 5. THE EDITION

Scholarly editions that include born-digital materials are still extremely rare, and none, to my knowledge, addresses the born-digital nature of the documents. An example is the recent Dürrenmatt *Stoffe* edition (<https://fd-stoffe-online.ch/>), which takes into account *computertexte* (see the table of contents of the *Archive* section). However, the user is given little information about the source of the text presented: while for manuscripts, typescripts and prints the facsimile and metadata provide information about the support, the writing tools, the layout and more, for the electronic documents there is no information about the software and system used, and the user is left to wonder what the original documents look like and how they were created.

The challenges faced by the Dürrenmatt editors are widespread. Paradoxically, nowadays it is common to access and incorporate into a scholarly edition the digital facsimile of a mediaeval source, but the same is extremely difficult for a born-digital object. In institutions archiving literary sources, access to born-digital materials is mostly limited to the internal reading-room, where they can be consulted on computers prepared to this end. Most of the time, no online access is allowed, therefore no possibility to retrieve it for displaying in a scholarly edition, as it would be the case for digital facsimiles thanks to the IIIF standard. This is true not only for literary sources: Lise Jaillant, the UK PI for four AHRC and international projects on artificial intelligence and born-digital archives, recently stated that “[e]ven when digital data is publicly available (as in the case of web archives), users often need to physically travel to repositories to consult web pages” [4]. When talking about access to born-digital sources, the debate between emulation and simulation, which is beyond the scope of this paper, is also relevant.

In our case, the scientific and technical choices concerning the establishment of a critical text and the presentation of the edition depend on the nature of the work and on the results of the preliminary steps discussed here above. Witnesses that contain very different texts, for instance, requires specific editorial choices and rendering solutions, both in the case of paper and of born-digital sources.

## 6. CONCLUSIONS

This paper presents preliminary reflections on the scholarly edition of born-digital sources and provides a concrete case study to discuss potentially widespread issues.

Problematic aspects that have been encountered in the case study are the absence of a detailed catalogue and the difficulty to retrieve original metadata. A distinctive feature of born-digital sources is that they do not contain traces of the genetic process, unless ad hoc mechanisms are employed by the author (e.g., track changes, keystroke logging) or by the researcher (digital forensic methods). The use of specific tools (reading devices, file profiling and digital forensic software) employed by archival institutions or by external researchers is indeed central in dealing with this kind of materials [9; 7 and previous contributions by the author]. When doing so, ethical and legal aspects, which are beyond the scope of this paper, should be cleared.

In the 2013 chapter for *The Cambridge Companion to Textual Scholarship*, Kirschenbaum and Reside asked “what can the textual scholar do to ensure that the born digital material of today remains legible and accessible to the scholar of the future?” For the authors, the answer lay in the advocacy role of textual scholars in library and archive settings, to ensure that both the content and the materiality are preserved “in *useful* as well as usable form” (italic in the source) [6; cf. 3]. This paper wishes to provide evidences and reflections for contributing to the dialogue between textual scholars and archival institutions, by highlighting practices and issues that are specific to the edition of born-digital literary sources. To overcome some of the current obstacles, the collaboration with archival institutions and, when possible, authors, remains fundamental.

## REFERENCES

- [1] Carbé, Emmanuela. 2018. “Digitale d’auteur. Un archivio possibile?” In *Archivi editoriali. Tra storia del testo e storia del libro*, edited by Virna Brigatti, Anna Lisa Cavazzuti, Elisa Marazzi, and Sara Sullam, 216–25. Milano: Unicopli.
- [2] 2023. *Digitale d’auteur. Macchine, Archivi, Letterature*. Firenze University Press.
- [3] Gooding, Paul, Jos Smith, and Justine Mann. 2019. “The Forensic Imagination: Interdisciplinary Approaches to Tracing Creativity in Writers’ Born-Digital Archives.” *Archives and Manuscripts* 47 (3): 374–90. <https://doi.org/10.1080/01576895.2019.1608837>
- [4] Jaillant, Lise. 2022. *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*. Bielefeld: transcript.

- [5] Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine. 2010. "Digital Forensics and Born-Digital Content in Cultural Heritage Collections" Washington, D.C.: Council on Library and Information Resources. <https://www.clir.org/pubs/reports/pub149/>
- [6] Kirschenbaum, Matthew G., and Doug Reside. 2013. "Tracking the Changes: Textual Scholarship and the Challenge of the Born Digital." In *The Cambridge Companion to Textual Scholarship*, edited by Neil Fraistat and Julia Flanders, 1st ed., 257–73. Cambridge University Press. <https://doi.org/10.1017/CCO9781139044073.013>
- [7] Kirschenbaum, Matthew G.. 2021. *Bitstreams: The Future of Digital Literary Heritage (Material Texts)*. Philadelphia: University of Pennsylvania Press.
- [8] Lebrave, Jean-Louis. 2011. "Computer forensics : la critique génétique et l'écriture numérique." *Genesis. Manuscrits – Recherche – Invention*, no. 33 (October): 137–47. <https://doi.org/10.4000/genesis.633>
- [9] Ries, Thorsten. 2018. "The Rationale of the Born-Digital Dossier Génétique: Digital Forensics and the Writing Process: With Examples from the Thomas Kling Archive." *Digital Scholarship in the Humanities* 33 (2): 391–424. <https://doi.org/10.1093/lilc/fqx049>.
- [10] Van Hulle, Dirk. 2019. "Towards a Comparative History of the Literary Draft." *Archiv für das Studium der neueren Sprachen und Literaturen*, no. 1 (May). <https://doi.org/10.37307/j.1866-5381.2019.01.02>
- [11] Weston, Paul Gabriele, Primo Baldini, Emmanuela Carbé, and Laura Pusterla. 2019. "Archivi digitali di persona PAD - Pavia Archivi Digitali e gli archivi degli scrittori." *DigItalia* 14 (1): 31–54.
- [12] Weston, Paul Gabriele, Emmanuela Carbé, and Primo Baldini. 2016. "Hold It All Together: A Case Study in Quality Control for Born-Digital Archiving." *Qualitative and Quantitative Methods in Libraries* 5 (3): 695–710.

# Per un atlante delle biblioteche medievali: il progetto Mirabile-Atlas e una nuova prospettiva sulla circolazione della cultura manoscritta

Cristina Ricciardi<sup>1</sup>, Luca Ughetti<sup>2</sup>

<sup>1</sup> SISMEL Firenze, Italia – lcg.ughetti@gmail.com

<sup>1</sup> SISMEL Firenze, Italia – cristina.ricciardi@unibo.it

## ABSTRACT

Il progetto internazionale Mirabile-Atlas, sostenuto dalla SISMEL e dalla ZKSchindler Foundation, ha come obiettivo la creazione di un atlante digitale delle biblioteche e dei centri di cultura dell'Europa medievale. L'intervento proposto descrive le fasi di sviluppo del progetto, le competenze digitali messe in campo e le scelte operate riguardo alle possibilità di rappresentazione. Dopo un'analisi del funzionamento generale, si passa in esame la strutturazione dei *database* e i criteri di efficacia e funzionalità che hanno condotto all'interfaccia utente. L'ampiezza dei dati a disposizione permette di considerare a fondo i presupposti delle *spatial representation* nelle *Digital Humanities* e di concepire la visualizzazione non solo come supporto alla comprensione della cultura, ma come un canale d'accesso a nuove prospettive di ricerca. L'atlante a tutti gli effetti rappresenta un nuovo approccio alla conservazione del patrimonio manoscritto che si affianca alle risorse della conservazione materiale e rinnova in una dimensione virtuale la circolazione e la fruibilità dei codici.

## PAROLE CHIAVE

Atlas; Codex Inventory; Manuscript; Medieval Library; Spatial Representation

## 1. INTRODUZIONE

L'atlante Mirabile-Atlas è il risultato di un progetto di collaborazione internazionale tra la SISMEL (*Società Italiana per lo Studio del Medioevo Latino*) e la Zeno Karl Schindler Foundation, con lo scopo di creare una prima rappresentazione di scala europea sulla circolazione dei manoscritti, delle opere e degli autori nel patrimonio delle biblioteche medievali. Il lavoro è concepito come prima risposta alla diffusione e circolazione dei codici e come punto di partenza per ricerche più specifiche. Per arrivare a questi risultati, l'atlante organizza più livelli di gestione della documentazione, ciascuno dei quali richiede delle strategie di intervento e richiama alle basi degli studi nelle *Digital Humanities*.

Una sfida si impone a chi ricostruisce il patrimonio di una biblioteca medievale: rendere in un'immagine definita un flusso di codici che attraversa luoghi e tempi. Le idee circolavano infatti con il loro supporto materiale, passando di mano in mano tra enti e possessori diversi. Anche quando i codici si trovavano all'interno di una biblioteca, compilare un inventario non era un'operazione immediata: alcuni manoscritti erano legati con una catena ai banchi per essere consultati, altri erano conservati in casse sotto custodia, altri erano temporaneamente prestati per la copia o rimanevano nelle celle per la lettura privata dei monaci. A distanza di secoli, ricomporre questa galassia di conservazione e di circolazione testuale è uno degli obiettivi principali nello studio della cultura medievale e della ricezione delle opere. Sono due le vie che possono essere percorse: lo studio diretto dei manoscritti e quello degli inventari redatti in epoca medievale.

L'umanistica digitale offre gli strumenti per entrare di nuovo nei luoghi che erano al centro della rete culturale e sociale nel Medioevo. Il progetto Mirabile-Atlas ricomponde in un atlante digitale la distribuzione e la consistenza della cultura materiale scritta tra il VI e il XIV secolo, consentendo agli utenti di percorrere il patrimonio bibliotecario in senso cronologico, di cercare la presenza di un'opera e di un autore, di esaminare la consistenza dei manoscritti e delle opere contenute. L'attenzione per la rappresentazione spaziale da parte dell'umanistica digitale è stata messa in relazione con lo *Spatial turn* che ha attraversato le discipline letterarie e le scienze sociali negli ultimi decenni. Nello studio *The Spatial Humanities*, che offre una prima importante panoramica dei progetti fino al 2010, si mette in rilievo come il focus disciplinare sui sistemi e tecniche GIS di rappresentazione geografica porti con sé in realtà un modo diverso di concepire l'approccio allo spazio (Bodenhamer et al. 2010). La possibilità di rappresentare lo spazio va oltre l'aspetto tecnico. Anche sul fronte didattico e accademico, nell'ultimo decennio questa applicazione nelle *Digital Humanities* si è andata espandendo e ha ricevuto riconoscimenti istituzionali, con la promozione di corsi di laurea dedicati e di laboratori di ricerca (Johanson 2012). L'anno successivo dalla sua fondazione, la rivista *Humanités numériques* dedica nel 2021 un numero alle «Humanités numériques spatialisées». Nel volume si susseguono interventi rivolti alle rappresentazioni spaziali di fonti

testuali in ambito letterario e nelle scienze sociali e dedicati all'analisi delle più ricorrenti problematiche di acquisizione dei riferimenti geografici passati e di impiego di questi dati nell'archeologia (Brando, Frontini et al. 2021). Tra gli ambiti di studio attraversati dai contributi, la cultura e la letteratura medievale sono scarsamente rappresentate, con un solo articolo che tratta dei siti archeologici di area adriatica dalla tarda antichità all'alto medioevo (Tassaux 2021).

Le risorse delle *Digital Humanities* applicate in età medievale devono confrontarsi con una documentazione ricca di informazioni che occorre calare nel contesto e nelle intenzioni di chi le ha prodotte, così che si impone una costante riflessione sullo sviluppo di appropriate metodologie storiografiche. Già nelle prime applicazioni si era sottolineata la necessità di tradurre gli strumenti informatici in risorse utili alla disciplina, realizzando «una sorta di *translatio studii*, con tutta l'inevitabile serie di nuovi interrogativi e timori che ciò comporta» (Bordini 2001). Gli strumenti di ricerca in ambito medievistico, e specificamente testuale, hanno permesso di concentrarsi sulle fonti di tipo documentale e manoscritto per l'accesso alla dimensione culturale, integrandosi così agli studi di codicologia, paleografia e diplomatica. Uno dei principali progetti che si fondano sulle risorse manoscritte è il portale *Biblissima*, che nella sezione relativa ai manoscritti dà la possibilità di visualizzare su una mappa i luoghi di origine e di attuale conservazione dei codici. In modo coerente con i suoi scopi, la rappresentazione spaziale rimane una risorsa accessoria e non dà la possibilità di integrazioni con altri strumenti di visualizzazione spaziale, come grafici di statistica o mappe topografiche. Ad oggi due progetti sostenuti dalla Sismel hanno dato come esito la creazione di una mappa digitale. Il primo è la “Carta interattiva della Toscana fino al 1325”, che rientra nelle celebrazioni per i settecento anni dalla morte di Dante Alighieri e si focalizza sui luoghi e sugli enti della regione presenti nella sua opera. Il secondo è l’“Atlante dei Luoghi della Cultura Scritta nella Toscana Medievale”, dedicato al censimento degli enti attivi in Toscana in epoca medievale e alla descrizione puntuale del relativo patrimonio manoscritto.

Queste ricerche rappresentano un punto di partenza metodologico per considerare le soluzioni impiegate e l'organizzazione dei dati. Il tutto viene adattato ad una dimensione fortemente originale. Il progetto Mirabile-Atlas infatti si smarca in modo rilevante dalla maggior parte dei *case study* per un'articolazione del progetto su più momenti, l'impiego di strumenti informatici tra loro integrati, per la consistente entità di dati elaborati e per il carattere *in progress* del lavoro, che prevede un costante ampliamento. Nelle sezioni successive viene descritto il funzionamento dell'atlante diviso nelle tre fasi di sviluppo parallelo. Per ciascuna fase si descrivono gli obiettivi e l'impianto teorico implicato, con le soluzioni attuate in risposta ad esigenze puntuali.

## 2. FUNZIONAMENTO E PROCESSI

L'atlante Mirabile-Atlas ad oggi è in fase di sviluppo e non è ancora accessibile all'utente finale. Si presenta come una pagina web che rende immediatamente disponibile la maschera di ricerca da cui impostare i parametri della *query* (fig. 1): l'utente può selezionare l'autore, l'ente che possedeva manoscritti e inventari medievali e può restringere la ricerca per cronologia. I parametri di ricerca possono essere incrociati. Il risultato è quindi visualizzato sull'atlante con puntatori georeferenziati per enti, città e per aree più estese, quali regioni e stati (fig. 2). I dati disponibili per l'atlante ad oggi sono localizzati soprattutto in Toscana, Umbria, Marche, Abruzzo, Molise, così che i risultati di ogni ricerca sono focalizzati principalmente in quest'area.

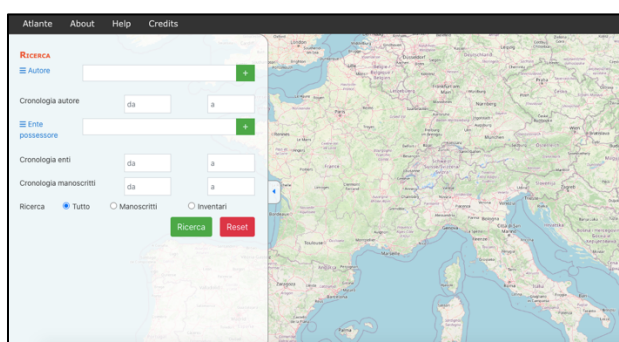


Figura 1. Pagina iniziale di Mirabile-Atlas.

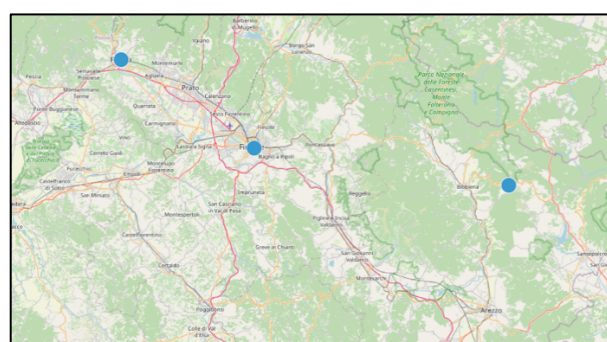


Figura 2. Zoom parziale dei risultati di ricerca per l'autore Guglielmo d'Auxerre, *Guillelmus Autissiodorensis*, organizzati per città.

L'atlante è l'esito di un articolato processo di sviluppo che riguarda sia i *database* sia l'interfaccia utente. A livello strutturale è costituito da due parti, una parte di *front-end* e una di *back-end*, entrambe realizzate in JavaScript. La ripartizione segue uno schema consolidato nel *web development* e risponde all'esigenza di dividere la parte del progetto

che si presenta allo *user* dalla parte che conduce la ricerca secondo i parametri richiesti e che, in modo circolare, fornisce al *front-end* i dati da visualizzare. La parte di *back-end* del progetto Mirabile-Atlas condivide la struttura generale con l'Archivio Integrato, che gestisce la maggior parte delle attività di ricerca della Sismel e della Fondazione Franceschini (FEF) all'interno di una collezione condivisa di dati, rimanendo indipendente nel suo funzionamento. Da parte del *server side* le ricerche sono effettuate utilizzando il motore di ricerca Elasticsearch e possono essere visualizzate attraverso l'interfaccia Kibana per uso interno, mentre sono restituite come file JSON per la visualizzazione sull'atlante digitale.

### 3. STRUTTURA DEI DATABASE

Il funzionamento dell'atlante si appoggia su due macro-categorie di dati impiegati: le schede inventariali e i dati dei singoli manoscritti. Questa divisione è rispecchiata nei due *database* di riferimento, denominati "Inventari" e "Manoscritti". La strutturazione di queste due collezioni deriva da una selezione di dati specifici, operata a partire dai risultati di ricerche già attive all'interno della Sismel. Fin dalle fondamenta il progetto è infatti concepito per essere trasversale rispetto a queste risorse: una parte consistente dei codici disponibili all'interrogazione proviene da progetti paralleli, tra cui CODEX, che si occupa del censimento e della catalogazione su base informatica del patrimonio manoscritto toscano. Sui dettagli di questo lavoro si rimanda a contributi specifici (*Codex Studies*; Pomaro 2020; Pomaro 2014; Marchiaro, Zamponi 2007).

Gli inventari delle biblioteche redatti in epoca medievale attestano la traccia tangibile di passaggi di codici e la disponibilità di letture che gravitava intorno ai centri di cultura. Questi elenchi permettono di ricostruire non soltanto la composizione e le successive vicende di fondi singolarmente localizzati, come si è anticipato, ma anche la rete di scambi che sussistevano tra le biblioteche, le persone e, di conseguenza, di formulare un'ipotesi sull'effettiva circolazione del sapere. L'attenzione per questa tipologia di fonte si è consolidata a fine XIX secolo con il lavoro del codicologo e bibliotecario Theodor Gottlieb, *Über Mittelalterliche Bibliotheken* (Leipzig, 1890), che si colloca a valle di una tradizione secolare di studi europei sulla raccolta ed edizione di documenti inventariali. Un'essenziale ricostruzione degli indirizzi di ricerca da qui scaturiti è di Giovanni Fiesoli ed Elena Somigli (2009, XII-XVI). Agli elenchi statici di carattere inventariale, Somigli affianca una vasta documentazione che rende l'idea del dinamismo proprio di un materiale altrimenti muto. Oltre dunque ai documenti redatti da istituzioni civili e religiose, si rileva tutta la documentazione relativa alla «produzione, alla fruizione e alla conservazione del libro, in cui sono compresi inventari, testamenti, donazioni, atti di compravendita, commissioni librerie, note di prestito e di pegno (Somigli 2013, XIII)». Gli agenti che partecipano di questi scambi non sono soltanto enti fisici, ma anche singole personalità che attestano il possesso, l'acquisto o la donazione di beni.

Ogni informazione relativa ai manoscritti e agli inventari, per essere utilizzabile dall'atlante, deve essere trasferita nel formato digitale che costituisce l'ossatura dei *database* "Manoscritti" e "Inventari". Relativamente ai manoscritti, questo passaggio viene effettuato attraverso una scheda catalografica digitale organizzata in diversi campi. Alcuni prevedono il completamento con un testo libero, per esempio nel caso del numero di fogli totali o delle dimensioni del codice, mentre altri campi attingono a liste esistenti, come nel caso del materiale, dei luoghi e degli enti legati al manoscritto. Il riconoscimento dell'ente possessore è decisivo nell'ottica della ricerca, perché attraverso l'appartenenza ad un preciso fondo l'atlante colloca il manoscritto nella mappa, riportandolo in più luoghi se il codice incontra trasferimenti in età medievale. La visualizzazione di queste informazioni permette dunque di tracciare i suoi spostamenti, partendo dall'origine e percorrendo le varie tappe cronologiche e geografiche.

Allo stesso modo, uno specifico *form* permette di inserire nel sistema i dati relativi agli inventari che vanno a costituire il *database* "Inventari". Questo è strutturato su due livelli successivi sul piano gerarchico. Il primo livello testimonia l'esistenza di un inventario medievale associato all'ente o alla persona fisica che possedevano gli elementi elencati. Il secondo livello è rappresentato dalle singole voci inventariali che sono contenute in un'oggetto del primo livello: le informazioni su opere e autori qui contenute sono state manualmente riconosciute e rese ricercabili per il progetto Mirabile-Atlas. Questi due livelli appunto sono di ordine gerarchico, per cui si ha la seguente struttura:

- Livello 1: Inventari, atti di acquisto, note di possesso, etc.
- Livello 2: Insieme delle singole voci contenute in un inventario, nota di acquisto etc., di cui sono riconosciuti nel progetto gli autori e le opere menzionate.

I due livelli compongono insieme il *database* inventariale e operativamente vengono accresciuti attraverso una bibliografia pertinente. La risorsa principale da cui sono tratti i dati primari sono le schede dell'archivio RICABIM (*Repertorio di Inventari e Cataloghi di Biblioteche Medievali*), pubblicate in diversi volumi su base regionale. Giovanni Fiesoli, partendo dal censimento cartaceo degli inventari e del materiale documentario dei fondi librari (Fiesoli et al. 2009), aveva strutturato una prima griglia di riferimento per la trasposizione online, che è stata adattata e amplificata dal progetto Mirabile-Atlas. Nello specifico, ciascuna scheda è organizzata per luogo (per città o per località) e analizza un documento che riporta



notizia della circolazione manoscritta, con relativa datazione. La raccolta era già in potenza concepita per essere poi importata in un archivio digitale, per cui la struttura di base costituisce un riferimento compatibile con il tipo di informazione necessaria all'atlante. Si è venuto così a creare un repertorio informatizzato che può essere interrogato attraverso i canali di ricerca definiti dai metadati delle singole schede. È possibile cercare tutta la documentazione legata per esempio a uno specifico convento, che abbia magari sviluppato al proprio interno un fondo librario variamente accresciuto, così come quella relativa a un giurista o un medico.

Il secondo livello è costituito dalle voci inventariali elencate all'interno del singolo documento di primo livello. Il riconoscimento dei codici e delle attribuzioni si realizza nel progetto facendo riferimento alla bibliografia specifica e alle edizioni degli inventari. Occasionalmente, una parte di queste è reperibile online su *Internet culturale* grazie al progetto "Gli Inventari delle Biblioteche Medievali italiane in rete (secc. IX-XVI)" che le ha rese accessibili in digitale. Il secondo livello è necessario per rilevare la presenza di un autore e, in prospettiva, per effettuare una ricerca delle singole opere nell'atlante digitale. Questa ricerca sarà presto eseguibile attraverso l'interfaccia dell'atlante nella forma che è già visibile in Mirabile, il sito che rende fruibili le informazioni dei singoli progetti della Sismel. Cercando una singola opera o un autore, è possibile vedere gli inventari medievali che ne danno notizia (fig. 3):

Assisi (Perugia) - Biblioteca papale. 160- 160. It. librum summe de fide, spe et caritate editum a fratre Thoma de Aquino cum post. et corio rubeo. Bo. 253. Jo. 65.  
 Assisi (Perugia) - Biblioteca papale. 30- 30. It. primam partem secunde partis libri fratris Thome de Aquino. 144. 149. Jo. 33. 34. 185.  
 Assisi (Perugia) - Biblioteca papale. 59- 59. It. librum (s.Thomae) questionis de malo: *utrum malum sit aliquid*, cum post. et corio viridi. Jo. 166.  
 Assisi (Perugia) - Matthaheus de Aquasparta cardinalis OFM. 20- [32a] Prima pars secunde partis summe Thome  
 Assisi (Perugia) - Matthaheus de Aquasparta cardinalis OFM. 42- [16b] Questiones Thome disputate  
 Assisi (Perugia) - Matthaheus de Aquasparta cardinalis OFM. 47- [21b] Prima pars summe fratris Thome  
 Assisi (Perugia) - Matthaheus de Aquasparta cardinalis OFM. 5- [17a] Postille Thome super Job  
 Assisi (Perugia) - Matthaheus de Aquasparta cardinalis OFM. 73- [47b] Postille super Iob

Figura 3. Zoom parziale degli inventari che danno notizia di un'opera di Tommaso d'Aquino.

I due *database* costituiscono un insieme di dati che si presta a rappresentazioni spaziali che vanno oltre la dimensione geografica. In questo senso, uno sviluppo del progetto Mirabile-Atlas prevede di integrare nella mappa anche la possibilità di rappresentare i dati attraverso grafici. Una possibilità di visualizzazione nasce per esempio dall'associazione di opere e autori ad una categoria del sapere medievale di riferimento, in modo tale da ottenere un grafico circolare che mostri la divisione in materie delle opere conservate da un ente possessore o in un'area geografica. Di seguito i due grafici che si ricavano considerando la regione Toscana e l'Umbria a confronto (fig. 4):

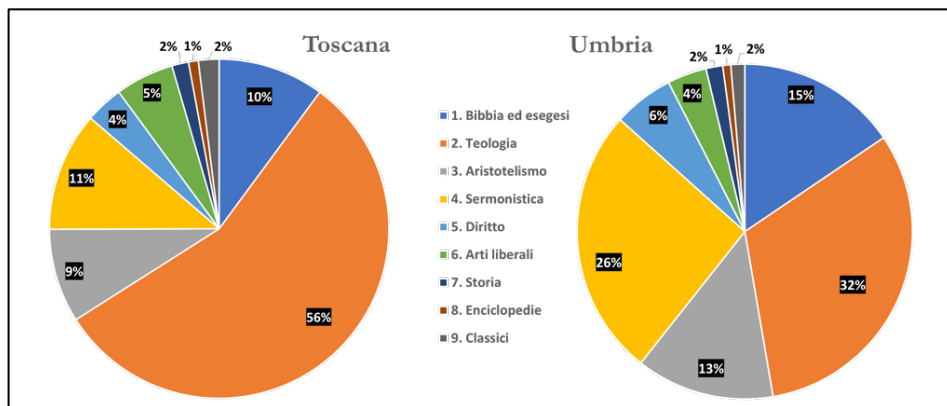


Figura 4. Distribuzione delle discipline contenute negli inventari degli enti censiti in Toscana e Umbria.

Ferma restando la necessaria semplificazione nell'associare opera e materia, lo strumento ha piena funzionalità quando risponde alla personalizzazione del singolo studioso che può scegliere la *ratio* propria dei raggruppamenti. I due grafici presentati sono una soluzione esemplificativa che parte dalla divisione di Edoardo d'Angelo in *La letteratura latina medievale: una storia per generi* (Roma, 2011), utilizzando l'interrogazione dei dati che al momento è disponibile in fase di sviluppo attraverso l'interfaccia Kibana. Reso accessibile all'utente finale, un simile strumento avrebbe lo scopo di affiancare la ricerca di opere ed autori all'interno di un fondo mostrando la distribuzione delle materie qui conservate. La sua implementazione rientra tra gli scopi a più lungo termine del progetto Mirabile-Atlas.

#### 4. INTERFACCIA UTENTE

L'interfaccia utente, attraverso cui avviene l'interrogazione della mappa, è concepita per essere di comprensione immediata: ogni campo esegue una ricerca semplice e specifica, mentre i criteri per una ricerca più elaborata nascono dalla combinazione di più campi di ricerca. La struttura di base è formata da due campi principali, "Autore" ed "Ente possessore",

a cui si aggiungono i campi per restringere la cronologia e per selezionare, sempre in ottica di ulteriore delimitazione, uno solo dei *database*. Per impostare la ricerca nel campo “Autore”, è possibile digitare direttamente il nome o selezionarlo da una lista di risultati fornita dalla codifica di BISLAM (*Bibliotheca Scriptorum Latinorum Medii Recentiorisque Aevi*), che segnala anche i dati cronologici e le varie forme in cui è attestato il nome dell’autore. Tramite il campo “Enti” possono essere selezionati anche i luoghi d’interesse: questo permette di condurre una ricerca più generale rispetto all’ente e soprattutto di individuare le singole persone, in quanto giuridicamente si possono ritenere ugualmente possessori.

Il base alla ricerca per “Autore”, i puntatori corrispondono ai luoghi legati ai due *database* da cui si attingono i dati. Fanno ovvero riferimento agli inventari (schede Ricabim) al cui interno sono presenti voci che identificano l’autore richiesto, o ai codici che testimoniano quell’autore. Cliccando sui singoli puntatori si apre una schermata con l’elenco degli enti collegati al luogo geolocalizzato sulla mappa e ciascuno riporta i propri risultati bipartiti in “Inventari” e “Manoscritti” (fig. 4). Questi ultimi conducono poi alle singole schede dedicate al materiale specifico in Mirabile, sfruttando il collegamento tra i progetti e i dati disponibili nell’Archivio Integrato. La bipartizione della schermata dei risultati tra i due *database* è stata impostata come scelta di default. All’interno della maschera di ricerca iniziale sono infatti presenti due sotto-campi per filtrare i risultati della ricerca: i campi “Inventari” e “Manoscritti”. La selezione è preimpostata su “Tutto”, per cui la ricerca è condotta su entrambi i *database*. Nel caso l’utente ne selezioni solo uno, la ricerca sarà effettuata in quell’ambito e coerentemente anche la schermata mostrerà solo la sezione interrogata.

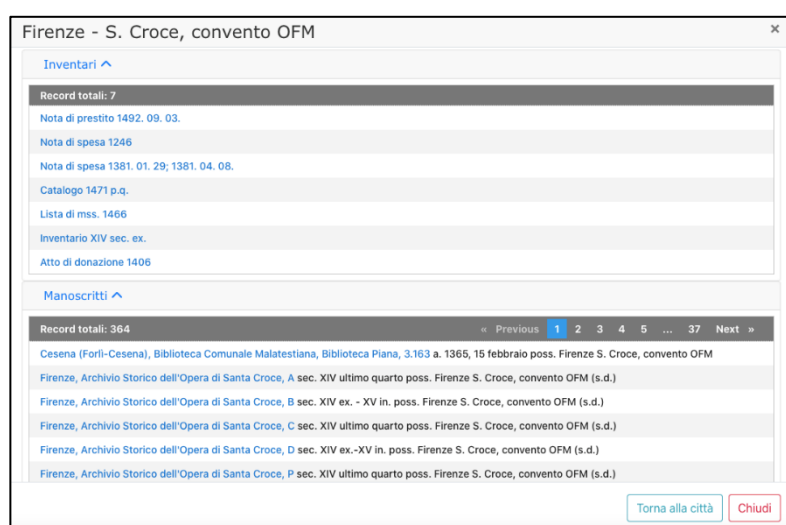


Figura 5. Finestra con l’elenco dei documenti inventariali e con i codici relativi all’ente selezionato, in questo caso Santa Croce di Firenze.

Come si vede dalla rappresentazione della pagina iniziale (fig. 1), la ricerca può essere condotta anche secondo un criterio cronologico, utilizzando gli appositi campi dedicati alla cronologia dell’autore, degli enti o dei manoscritti. Nel caso dell’autore, l’arco cronologico impostato prevede di individuare gli autori vissuti in quel tempo. Per gli enti, si fa riferimento alla datazione attribuita ad essi nell’Archivio Integrato, mentre per i manoscritti alla datazione di composizione. Attraverso la combinazione delle risorse messe a disposizione, l’atlante digitale prodotto dal progetto Mirabile-Atlas ha il proposito di fornire agli studiosi uno strumento di partenza per itinerari di ricerca. L’impiego più immediato va a vantaggio degli studi archivistici e di storia delle biblioteche, ma la possibilità di ricostruire lo spostamento dei codici risponde a interrogativi di carattere codicologico e paleografico in modo altrettanto esauriente. La ricerca per autori, incrociando ulteriormente le informazioni fornite dagli inventari, costituisce poi un primo passo verso la ricostruzione della ricezione delle opere. Permette in definitiva di tratteggiare i contorni di una storia della mentalità dell’intellettuale medievale. In una prospettiva di recupero del patrimonio culturale, la costituzione di un archivio e di un sistema di rappresentazione digitale costituisce la controparte della conservazione materiale. Si pone infatti come uno strumento che facilita l’accesso al codice medievale e che ne può orientare l’analisi, in un contesto accademico e di fruizione più generalizzata.

## 5. RINGRAZIAMENTI

Il progetto Mirabile-Atlas è reso possibile dal sostegno della Sismel e della Zeno Karl Schindler Foundation, a cui va il ringraziamento degli autori. Al prof. Agostino Paravicini Bagliani, presidente della Sismel, e al prof. Jean-Yves Tilliette, Université de Genève e ZKS Foundation, decisivi per la nascita e la sua evoluzione. Alla dott.ssa Lucia Pinelli per il

fondamentale coordinamento scientifico e al prof. Francesco Santi, direttore della Sismel, per i preziosi consigli sullo sviluppo.

## BIBLIOGRAFIA

- [1] Atlante dei Luoghi della Cultura Scritta nella Toscana Medievale: <https://atlas.mirabileweb.it/toscana/atlas>
- [2] Biblissima: <https://portail.biblissima.fr/fr/>
- [3] Bodenhamer, David J., John Corrigan e Trevor M. Harris. The Spatial Humanities. GIS and the Future of Humanities Scholarship. Bloomington et Indianapolis: Indiana University Press, 2010
- [4] Bordini, Simone. “Medioevo in rete tra ricerca e didattica”. Quaderni storici 36 (2001): 635-642.
- [5] Brando, Carmen, Francesca Frontini, Dominic Moreau, Mathieu Roche e Éric Masson, “Introduction. Humanités numériques et analyses spatiales: enjeux et perspectives” Humanités numériques 3 (2021): 1-28. <https://journals.openedition.org/revuehn/2038>
- [6] Codex Studies, rivista: <https://www.sismelfirenze.it/index.php/riviste/codex>
- [7] D’Angelo, Edoardo. La letteratura latina medievale: una storia per generi. Roma: Viella, 2011.
- [8] Elasticsearch: <https://www.elastic.co/elasticsearch/>
- [9] Fiesoli, Giovanni e Elena Somigli. RICABIM 1. Italia. Toscana. Firenze: SISMEL Edizioni del Galluzzo, 2009
- [10] I luoghi della cultura in Toscana fino al 1325: <https://atlas.mirabileweb.it/atlas>
- [11] Internet Culturale, sezione Inventari delle biblioteche medievali italiane: <https://www.internetculturale.it/it/41/collezioni-digitali/26175/>
- [12] Johanson, Chris, Elaine Sullivan, Janice Reiff et al. “5. Teaching Digital Humanities through Digital Cultural Mapping” in Digital Humanities Pedagogy: Practices, Principles and Politics, a cura di Brett D. Hirsch, 121-49. Cambridge: Open Book Publishers, 2012. <https://books.openedition.org/obp/1626>
- [13] Kibana: <https://www.elastic.co/kibana/>
- [14] Marchiaro, Michaelangiola, e Stefano Zamponi. Conoscere il manoscritto: esperienze, progetti, problemi. Dieci anni del progetto Codex in Toscana. Atti del Convegno internazionale, Firenze, 29-30 giugno 2006. Firenze: SISMEL Edizioni del Galluzzo, 2007
- [15] Pomaro, Gabriella. In margine al progetto Codex. Aspetti di produzione e conservazione del patrimonio manoscritto in Toscana. Pisa: Pacini Editori, 2014.
- [16] Pomaro, Gabriella. Attorno a Codex. Nuovi materiali e approfondimenti. Firenze: SISMEL Edizioni del Galluzzo, 2020
- [17] Somigli, Elena. RICABIM 3. Italia. Umbria, Marche, Abruzzo, Molise. Firenze: SISMEL Edizioni del Galluzzo, 2013
- [18] Tassaux, Francis. “AdriAtlas et IllyrAtlas, deux atlas informatisés de l’Antiquité romaine et du haut Moyen Âge” Humanités numériques 3 (2021): 1-31. <https://journals.openedition.org/revuehn/1543>

# Dai metadati ai dati, dai contesti ai contenuti: aumentare la descrizione archivistica

Federico Valacchi<sup>1</sup>

<sup>1</sup> Università degli Studi di Macerata, Italia – federico.valacchi@unimc.it

## ABSTRACT

L'intervento si propone di valutare in chiave comparativa la dimensione culturale, tecnica e scientifica degli strumenti archivistici per individuare tecniche euristiche che agiscano sui contenuti e sui dati e vadano oltre i fisiologici limiti quantitativi della ricerca archivistica che si manifestano anche negli archivi digitali o digitalizzati

## PAROLE CHIAVE

Descrizione archivistica; archivi digitali; collezioni digitali; learning machine; strumenti di ricerca

## 1. INTRODUZIONE

La descrizione archivistica, per lunga e inevitabile tradizione, procede muovendo dallo studio dei contesti verso un'individuazione necessariamente approssimativa dei contenuti<sup>1</sup>. La buona volontà della mediazione, nella maggior parte dei casi, si deve infatti confrontare con una brutale dimensione quantitativa. Gli archivi sono oceani di informazione e gli strumenti e gli approcci che potremmo definire tradizionali fanno oggettivamente fatica a dare istruzioni davvero puntuali ai naviganti. D'altra parte è altrettanto inevitabile mantenere i contenuti agganciati ai contesti, se vogliamo continuare a parlare di archivi in senso proprio, per quanto allargati.

Dal punto di vista squisitamente metodologico si pone quindi il problema di andare oltre il canone descrittivo senza tradire le coordinate di fondo della descrizione archivistica e del metodo stesso.

*Tempus fugit* e anche le nostre idee di descrizione e, soprattutto di fruizione, devono fare i conti con la voracità dell'obsolescenza innescata dalle violente e pervasive accelerazioni tecnologiche cui la nostra società nel suo insieme è sottoposta. In particolare, quello che potremmo definire un costume tecnologico indotto ha diversificato l'utenza degli archivi, rendendola più esigente<sup>2</sup>. Siamo ormai abituati ad ottenere risposte piuttosto che a porci domande. Ci muoviamo dentro alle logiche per certi aspetti perverse dei motori di ricerca. Molti utenti faticano perciò a comprendere come proprio gli archivi, luoghi deputati alla custodia e all'uso dell'informazione, stentino a rispondere in maniera puntuale alle loro interrogazioni. Come abbiamo detto, esistono ragioni incontestabili per spiegare questa approssimazione informativa, ma limitarsi a giustificarla non basta più.

“La domanda di risposte” non può più essere ignorata, anche alla luce di una tecnologia che amplifica i bisogni e sembra suggerire soluzioni allettanti.

Il problema è reale e complicato. Mette in gioco le tecnologie, anche venturose, e le politiche di digitalizzazione, ma tira in ballo anche la nostra idea di mediazione. Le risposte che cerchiamo possono infatti incidere sulla descrizione archivistica, magari *aumentandola* per insegnare alla artificiale intelligenza delle macchine a districarsi tra le parole dei documenti.

Si potrebbe obiettare che c'è il rischio di snaturare la disciplina, ma in realtà l'archivistica insegue da sempre la mutevolezza degli archivi e dei bisogni che essi devono soddisfare. Si può imparare senza abdicare, ci si possono porre obiettivi ambiziosi senza rinunciare ai propri metodi e ai propri strumenti. Se

---

<sup>1</sup> Antonio Romiti, «I mezzi di corredo archivistici e i problemi dell'accesso», *Archivi per la storia* III, fasc. 2 (1990): 217–46.

<sup>2</sup> Barbara Lazenby Craig, «Old Myths in New Clothes: Expectations of Archives Users», *Archivaria* 45, (1998): 118–26.

il formato e la fisionomia dell'inventario archivistico potranno modificarsi e diventare qualcosa di diverso dalla percezione che ne abbiamo sempre avuto, non necessariamente verranno meno le ragioni profonde per cui si producono strumenti di ricerca.

Bisogna chiedersi se sia possibile aumentare la portata informativa degli strumenti archivistici in senso ampio, salvaguardando i sistemi di relazioni che governano ogni fondo archivistico, ma permettendo al tempo stesso agli utenti di estrarre più agevolmente il succo informativo dei singoli documenti, il cui reperimento giustifica l'intero lavoro archivistico.

Le tecnologie di cui disponiamo in misura apparentemente inesauribile possono darci indicazioni e perfino risposte a questo riguardo, ma dobbiamo essere disponibili a metterci in gioco, ferma restando la conoscenza di base di ogni fondo archivistico: l'archivio va prima descritto e poi riordinato, altrimenti ogni strategia empirica e tecnologica sarà vana se non controproducente.

Se l'archivio è ordinato, come abbiamo detto, potremmo cercare la soluzione in un concetto aumentato di descrizione archivistica, all'interno del quale ricondurre intanto processi di indicizzazione e/o di trascrizione selettiva dei documenti digitalizzati, finalizzata ad assecondare l'allenamento delle macchine nel riconoscimento automatico del testo. L'intento è quello di svegliare dal suo torpore l'immagine digitale, mummificata nel formato di copia fotografica dell'originale analogico.

Prima che tecnica e tecnologica la questione è metodologica e, in un certo senso, perfino antropologica. Occorre intanto rivisitare acquisizioni consolidate, per accettare soprattutto che *soggettazione* o *materie* possono non essere bestemmie archivistiche se le si declina nel modo opportuno e nella dovuta armonia tecnologica.

Sembra poi evidente che in questi ipotetici scenari l'archivistica non basta più a sé stessa. Si impongono nuove e costruttive alleanze con le altre discipline dell'universo documentario, allargando lo spettro della collaborazione anche ai diversi domini di gestione dell'informazione in senso stretto. Si va dalla paleografia alle digital humanities nel senso più ampio e nobile dell'espressione e si va oltre le pur sacrosante esigenze euristiche. Quella che si annuncia infatti è una battaglia della conoscenza contro la (dis)informazione digitale, che mette in gioco tutte le discipline di area LIS e, più in generale, suggerisce un confronto franco e concreto con le digital humanities, in cerca di comuni spazi di ulteriore sviluppo<sup>3</sup>.

## 2. METODO, STRUMENTI E RICERCA

Ogni fondo archivistico si manifesta solo nel suo riuso informativo nel tempo. Sono le sollecitazioni esterne a determinarne il valore specifico e per questa ragione il processo di mediazione archivistica accompagna l'intero ciclo vitale e ad esso deve adeguarsi. Gli strumenti di ricerca, ed in particolare gli inventari, sono essi stessi prodotti in divenire e sono prima di tutto testimonianze peculiari del clima culturale e scientifico da cui scaturiscono. L'inventario, dal punto di vista dell'uso, è esso stesso parte del contesto che mira a ricostruire. La sua necessaria approssimazione lo apre paradossalmente a possibili implementazioni anche successive alla "pubblicazione", soprattutto quando ci si riferisca a banche dati di descrizioni archivistiche. La ricerca che muove dalla ricerca asseconda un'economia circolare delle informazioni che, almeno in parte, può supplire ai limiti fisiologici cui abbiamo accennato<sup>4</sup>. Gli strumenti non sono astrazioni euristiche. Sono piuttosto il complesso risultato di un sistema di costruzione della conoscenza che affonda le sue radici in un metodo che nella sua essenza continua a funzionare, anche perché oggettivamente è l'unico che abbiamo. L'insieme delle risorse sempre più raffinate che la mediazione archivistica ha reso disponibili nel tempo ha dato sicuramente risultati più che soddisfacenti. La riflessione mai interrotta sulla natura e le finalità della descrizione archivistica ne è la tangibile dimostrazione e garanzia. Il metodo è l'impianto sintattico e grammaticale della lingua con cui si esprimono gli archivi. E ci protegge dal rischio concreto di un'ingovernabile anarchia documentaria.

---

<sup>3</sup> Marilena Daquino e Francesca Tomasi, «Digital Humanities e Library and Information Science. Attraverso le lenti dell'organizzazione della conoscenza», *Bibliothecae.it* 5, fasc. 1 (2016): 130–50, <https://doi.org/10.6092/ISSN.2283-9364/6109>

<sup>4</sup> Stefano Gardini, «Economie circolari dell'archivio: la carte di utenti e studiosi come archivi derivati», *Nuovi Annali della Scuola speciale per archivisti e bibliotecari* XXXV (2021): 237–77.

Il metodo storico, “archivistico” per definizione, continua a fare il suo lavoro, assecondato dagli standard di descrizione di prima e seconda generazione, ISAD(G) e RiC in testa, su cui avremo modo di tornare brevemente più avanti. Il risultato finale dei processi descrittivi continua a manifestarsi in una gamma di strumenti di ricerca che in molti casi sono costretti a puntare al contesto più che al contenuto. La massiccia digitalizzazione delle risorse archivistiche nel suo insieme, di fatto, non ha permesso di superare questo fondamentale limite, dal momento che si è fin qui limitata a riproporre strumenti e metodi consolidati, rivisitandoli alla luce di una maggiore potenza di calcolo. Il risultato della ricerca *on demand*, per chiamarla così, continua ad essere più una speranza che una certezza.

Abbiamo il metodo e abbiamo gli strumenti ma né l’uno né gli altri sono incisi sulla pietra<sup>5</sup>. In ragione di quella duttilità che si richiamava sopra, il metodo e gli strumenti che ne scaturiscono funzionano quando assecondano le ragioni effettive della produzione e della fruizione. È vero che ad un livello squisitamente funzionale le tecniche e le tecnologie con cui si costruisce il sistema di mediazione ne possono modificare nel tempo la configurazione, migliorandone anche il rendimento. Sarebbe però fuorviante confidare in un’evoluzione di taglio banalmente tecnologico. Indipendentemente dalle sue potenzialità e dalle sue caratteristiche strutturali, ogni inventario ha sempre risposto innanzitutto a problemi di organizzazione e restituzione di particolari famiglie di metadati. Un inventario è, appunto, un sistema *strutturato* di dati sui dati e tale resta indipendentemente da come lo si costruisce e restituisce.

Il metodo quindi ci serve ancora, anche se quando ci affacciamo sul mercato polimorfismo contemporaneo, e in particolare sugli archivi digitali, qualcosa può cambiare.

Il cambiamento più marcato e gravido di conseguenze non è solamente di natura meccanica e legato ai mezzi di produzione e alla natura dei supporti. Tendono infatti a modificarsi soprattutto le logiche di produzione e uso da cui poi deriva la diversificazione dei mezzi e degli strumenti.

Per capire meglio conviene innanzitutto definire quale sia *l’archivio* di cui parliamo, precisando che qui ci riferiamo ad aggregazioni digitali, sia native che generate a partire da processi di acquisizione di fondi archivistici analogici. Si può allora introdurre intanto una distinzione di massima, per quanto grossolana, tra due fondamentali tipologie. Da un lato stanno gli archivi informatici, tra i quali per estensione si possono considerare anche i siti web, con i problemi conservativi che pongono, e dall’altro la mole crescente di archivi o di porzioni di archivio digitalizzati a partire da consolidate sedimentazioni analogiche.

Nel primo caso vanno innanzitutto segnalate trasformazioni che già incidono in profondità sull’auspicabile “storicizzazione” di questi complessi documentari, sia in termini di processi conservativi che di adeguata contestualizzazione. L’impatto metodologico è forte, perché si mette in discussione la rassicurante univocità del *creator*. Si modifica il flusso funzionale della produzione, che non scaturisce più in maniera univoca da un solo soggetto, magari fortemente strutturato. La filiera documentaria tende a diluirsi, inseguendo le esigenze e le lusinghe di un’interoperabilità che non è solo linguaggio di scambio tra le macchine ma modo di agire di buona parte dei soggetti produttori.

La stessa conservazione, poi, perde consistenza e tracciabilità nei meandri di una delocalizzazione fisica che distrugge l’idea stessa di policentrismo e pone problemi di percezione unitaria ed univoca delle *universitas rerum* documentarie. L’archivio c’è, ma non si vede, e governarlo correttamente è più complicato. Si pone insomma la questione di un effettivo approccio storico e culturale a questi archivi. Non basta più manifestare la volontà di difendere la memoria digitale dall’obsolescenza o limitarsi a pensare a strategie di sopravvivenza che garantiscano la *long time preservation* degli oggetti digitali, siano esse migrazioni, cloud o blockchain. Il passo da fare sembra essere quello di porre la questione al giusto livello politico e culturale, riflettendo seriamente su un modello conservativo nuovo e adeguato al presente e al futuro della produzione documentaria.

Capire *come* conservare è vitale ed è la nuova urgente configurazione dell’idea di base di tutela. Se è vero che si conserva per consultare, è urgente riflettere anche sul *perché*.

La conservazione di lungo periodo costruisce *archivi storici in senso proprio*, più o meno in potenza, più o meno fruibili nell’immediato, ma degni delle particolari attenzioni che da sempre riserviamo a questa fase del ciclo vitale. La dimensione storica e culturale dell’accesso va perciò considerata parte integrante del

---

<sup>5</sup> Federico Valacchi, «Quiddam divinum. Riflessioni sul metodo storico», *Archivi XV*, fasc. 1 (2020): 69–87.

lavoro di progettazione, evitando di correre il rischio di affidare il recupero delle informazioni a impalpabili sistemi di information retrieval che, per quanto potranno imparare, rischiano di restare troppo generici ed evasivi.

Se ammettiamo che i progressi del metodo possano riuscire a metabolizzare queste trasformazioni, negli archivi digitali nativi il problema del recupero degli atomi informativi tecnicamente non si pone perché in presenza di documenti digitali sarà sempre possibile operare una ricerca full text. A patto naturalmente che l'archivio sia costruito in modo tale da garantire sempre la contestualizzazione del dato e che sia appunto un archivio, non un grande contenitore di bit in ordine sparso. In questi archivi l'inventario, sempre ammesso che lo vogliamo chiamare ancora così, non è più *dell'archivio* ma *nell'archivio*, ne fa parte integrante come peculiare funzionalità di ricerca, con tutto ciò che ne consegue anche a livello di tassonomia e di definizioni degli strumenti nel loro insieme.

Diverso il caso delle acquisizioni di documenti o nuclei di documenti provenienti dall'enorme eredità analogica, cioè della cosiddetta digitalizzazione delle fonti primarie.

La prima questione da affrontare al riguardo è quella, decisiva, della contestualizzazione/ricontestualizzazione, cioè del rapporto tra l'evidenza digitalizzata e l'integrità del fondo originario a cui si attinge. In presenza di processi di digitalizzazione selettiva l'ansia di costruire grandi serbatoi di "cose" digitali, dove la quantità vince sui sistemi di relazioni, può risolversi in rigenerazioni informative, se non in vere e proprie degenerazioni, le cui conseguenze possono essere piuttosto serie.

La dematerializzazione non si può negare, è semplicemente un dato di fatto. Si può però tentare di interpretarla, se non di governarla. Nel pieno di anni ipermnemonici, l'archivistica può dare il suo contributo a una più generale riflessione critica su quello che facciamo con le nostre tecnologie. Non si deve inventare nulla, basta seguire secolari processi virtuosi. L'ordine e l'inventario continuano ad essere i veri garanti di una coscienza critica dell'archivio, indispensabile soprattutto negli sviluppi non sempre coerenti della digitalizzazione. Nelle politiche dematerializzanti, quindi, prima dovrebbero arrivare gli inventari e poi gli *oggetti* che essi descrivono o introducono. Una digitalizzazione *object oriented*, senza adeguata descrizione e senza ordinamento preventivo, è un'anatra zoppa.

A prescindere da ogni altra considerazione, e dalla sua intrinseca qualità, nessuno degli strumenti attuali sembra però avere ad oggi la forza di superare i limiti non scritti della mediazione archivistica. Sono molto ma non tutto, si può fare di più.

Bisogna allora che il meccanismo faccia uno scatto, e si entri nel merito del recupero del dato all'interno del singolo oggetto digitale. Dobbiamo cioè sforzarci di spostare l'attenzione dai metadati ai dati, rilanciando, almeno in prima battuta, alcune strategie proprie del metodo per materia e basate in sostanza su particolari marcature del testo. L'indicizzazione e la soggettazione supportano già l'efficacia della gestione documentale e sostengono le attività di classificazione nella fase corrente. Queste tecniche possono dare il loro contributo anche nei fondi storici, soprattutto quando le si usi con la dovuta prudenza<sup>6</sup>.

La usuale descrizione archivistica può essere incrementata da ulteriori metadazioni, magari in forma di tag. Un tag, in questo senso, è un metadato che avvicina al contenuto, per quanto sia anch'esso una forma di interpretazione di chi lo genera.

Ci si può chiedere poi se esistano altre forme possibili di riconoscimento del testo e se si possa quindi spostare l'azione euristica dai metadati ai dati. Il problema di base è quello, noto da molto tempo, della difficoltà che una macchina incontra nel riconoscere nei segni dei significati lungo il processo di *handwritten text recognition*. Nello specifico, le tecnologie HTR nella loro costante evoluzione sembrano promettere risultati di sicuro interesse<sup>7</sup>.

---

<sup>6</sup> Roberto Guarasci e Mauro Guerrini, *Cos'è l'indicizzazione* (Milano: Editrice Bibliografica, 2022).

<sup>7</sup> Denis Coquenot, Clement Chatelain, e Thierry Paquet, «Handwritten text lines to whole documents», in *ORASIS 2021* (Saint Ferréol (France): Centre National de la Recherche Scientifique [CNRS], 2021), <https://hal.science/hal-03339648>



### 3. CONCLUSIONI

Una delle suggestioni più forti che i processi di dematerializzazione suscitano nel dominio degli archivi, e degli archivi storici in particolare, prende forma nella speranza di riuscire a disporre di documenti nei quali sia possibile operare puntuali ricerche per parola.

Le esorbitanti quantità informative con cui ci si confronta e l'indomabile anarchia della parola scritta rimangono però ostacoli di tutto riguardo. Ad oggi gli automatismi di ricerca passano ancora da un assiduo lavoro di trascrizione "manuale", finalizzato ad accrescere l'esperienza cognitiva della macchina per allenarla al riconoscimento dei segni. Trascrizione e verifica dei risultati dell'apprendimento sono gli strumenti di un lavoro tanto più moderno quanto antico e multidisciplinare.

Siamo a tutti gli effetti nel quadro di una descrizione archivistica aumentata, che non si ferma all'identificazione dell'oggetto, ma cerca di coglierne anche il contenuto, passando appunto da sistema di metadati a uno di dati contestualizzati. La descrizione identifica l'oggetto e la trascrizione lo svela, in un crescendo che sappia offrire al software materiale di confronto in grado di "allenarlo" e di potenziarne le performances cognitive specifiche.

Una descrizione che punti ai contenuti oltretutto ai contesti si arricchisce delle trascrizioni di porzioni selezionate del fondo, affidandosi a una logica incrementale, sorretta dal learning machine e dalla sua possibile crescita cognitiva specifica.

Si pone certamente il problema della selezione e dell'alterazione dei vincoli costitutivi, con tutti i rischi che ne conseguono. In linea teorica la selezione contraddice il mito dell'avalutatività, ma lo stesso metodo che sembra porre dei limiti può risolvere la contraddizione. Se l'ordine conferito al fondo e il suo inventario ci tutelano, e se le finalità dell'azione sono esplicitate, niente proibisce infatti di pensare a approfondimenti "tematici" su porzioni del fondo. Si potrebbe cioè immaginare una postproduzione degli strumenti, arricchita da opzioni di ricerca capaci di spingersi in profondità non tanto nelle relazioni, ma nelle parole di cui ogni fondo alla fine è costituito. La tematizzazione non è reato, se è sostenuta da un'effettiva contestualizzazione e muove magari dall'analisi dei bisogni prioritari della ricerca che è possibile stabilire a partire dai comportamenti degli utenti. Passare da una generica iconografia digitale, fatta di immagini inerti, a una restituzione dinamica dei contenuti dovrebbe anzi essere uno degli obiettivi prioritari di una digitalizzazione virtuosa dei documenti di archivio.

Se volessimo recuperare il linguaggio degli standard, si tratta di passare dal potere incontrastato della multivellarità relazionale di ISAD(G) alla rete multidimensionale di significati e di rinvii logici e semantici di RiC<sup>8</sup>. Non è del resto un caso che proprio RiC, standard di seconda generazione, dichiari esplicitamente la sua dipendenza anche dalla tecnologia di cui disponiamo, cioè da una tecnologia sempre più duttile, potente e capace di sciogliere nodi atavici che arrivano direttamente dalla bidimensionalità descrittiva analogica.

Sono naturalmente ipotesi e soluzioni da verificare con tutte le precauzioni del caso ma che promettono sviluppi interessanti, che vanno anche oltre il recupero del dato secco.

Questo processo di crescita incrementale della conoscenza potrebbe in prospettiva fare affidamento anche sul dato quantitativo garantito dalle ricerche che si sviluppano dalle ricerche. L'approssimazione informativa si può infatti combattere non solo grazie all'intelligenza artificiale ma anche a quella "collettiva", tutta umana ed esperienziale.

In definitiva, quindi, prima ancora di immaginare le soluzioni tecnologiche, occorre abbracciare le logiche di una descrizione integrata, finalizzata alla costruzione dei sistemi interculturali che si affacciano anche dalle pagine di RiC.

---

<sup>8</sup> Giorgia Di Marcantonio, «Resource Description and Access e il modello concettuale Records in Contexts. A Conceptual Model for Archival Description: oggetti comparabili?», *JLIS.it* 9, fasc. 1 (2018): 128–35, <https://doi.org/10.4403/jlis.it-12412>; Pierluigi Feliciati, «Archives in a Graph. The Records in Contexts Ontology within the framework of standards and practices of Archival Description», *JLIS.it* 12, fasc. 1 (2021): 92–101, <https://doi.org/10.4403/jlis.it-12675>

Si tratta di capire se e in che modo l'esperienza archivistica riesca a dialogare con le altre discipline dell'informazione e sostenere l'apprendimento delle macchine, per aprire i documenti archivistici alle potenzialità di una ricerca puntuale e quanto possibile indipendente dalle rigidità gerarchiche.

In palio c'è la possibilità di svincolare i fondi archivistici dalla figura archetipica del soggetto produttore, per farli confluire dentro a quadri informativi più ampi e articolati, dove il confronto con le digital humanities si arricchisce di ragioni e di speranze descrittive.<sup>9</sup> E dove l'archivistica può recuperare un ruolo importante in contesti del tutto mutati ma non per questo da trascurare.

## BIBLIOGRAFIA

[1] Coquenet, Denis, Clement Chatelain, e Thierry Paquet. «Handwritten text recognition: from isolated text lines to whole documents». In ORASIS 2021. Saint Ferréol (France): Centre National de la Recherche Scientifique [CNRS], 2021. <https://hal.science/hal-03339648>.

[2] Craig, Barbara Lazenby. «Old Myths in New Clothes: Expectations of Archives Users». *Archivaria* 45, fasc. January (1998): 118–26.

[3] Daquino, Marilena, e Francesca Tomasi. «Digital Humanities e Library and Information Science. Attraverso le lenti dell'organizzazione della conoscenza». *Bibliothecae.it* 5, fasc. 1 (2016): 130–50. <https://doi.org/10.6092/ISSN.2283-9364/6109>.

[4] Di Marcantonio, Giorgia. «Resource Description and Access e il modello concettuale Records in Contexts. A Conceptual Model for Archival Description: oggetti comparabili?». *JLIS.it* 9, fasc. 1 (2018): 128–35. <https://doi.org/10.4403/jlis.it-12412>.

[5] Feliciati, Pierluigi. «Archives in a Graph. The Records in Contexts Ontology within the framework of standards and practices of Archival Description». *JLIS.it* 12, fasc. 1 (2021): 92–101. <https://doi.org/10.4403/jlis.it-12675>.

[6] Gardini, Stefano. «Economie circolari dell'archivio: la carte di utenti e studiosi come archivi derivati». *Nuovi Annali della Scuola speciale per archivisti e bibliotecari XXXV* (2021): 237–77.

[7] Guarasci, Roberto, e Mauro Guerrini. *Cos'è l'indicizzazione*. Milano: Editrice Bibliografica, 2022.

[8] Romiti, Antonio. «I mezzi di corredo archivistici e i problemi dell'accesso». *Archivi per la storia* III, fasc. 2 (1990): 217–46.

[9] Tomasi, Francesca, e Marilena Daquino. «Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare». *JLIS.it* 6, fasc. 3 (2015): 13–38.

[10] Valacchi, Federico. «Quiddam divinum. Riflessioni sul metodo storico». *Archivi* XV, fasc. 1 (2020): 69–87.

---

<sup>9</sup> Francesca Tomasi e Marilena Daquino, «Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare», *JLIS.it* 6, fasc. 3 (2015): 13–38.

# ITINERA: progetto, database e prospettive

Nicola De Nisco<sup>1</sup>, Sandra Gorla<sup>2</sup>, Alessia Valenti<sup>3</sup>

<sup>1</sup> Università per Stranieri di Perugia, Italia – nicola.denisco@unistrapg.it

<sup>2</sup> Università degli Studi di Napoli Federico II, Italia – sandra.gorla@unina.it

<sup>3</sup> Università degli Studi di Siena, Italia – alessia.valenti2@unisi.it

## ABSTRACT

Il poster presenta la banca dati del progetto PRIN Itinera – *Italian Trecento Intellectual Network and European Renaissance Advent*, illustrando in particolare i modelli sviluppati per la descrizione dei *corpora* (testi, persone e manoscritti) e le scelte metodologiche e operative che hanno condotto all’elaborazione dello strumento digitale e degli altri *output*.

## PAROLE CHIAVE

Database; modellizzazione; LOD; web semantico; reti intellettuali

## 1. IL PROGETTO

ITINERA è un progetto diretto da Natascia Tonelli, responsabile dell’unità di Siena, a cui partecipano altre tre unità: Università Roma Tre, Università per Stranieri di Perugia e Università Federico II di Napoli. Il progetto è dedicato allo studio e all’analisi delle reti intellettuali individuabili a partire dalle corrispondenze di Petrarca. Attraverso il censimento e lo studio delle corrispondenze, ITINERA intende ricostruire l’universo culturale e storico del Trecento italiano ed europeo, indagando inoltre la prima fortuna delle opere petrarchesche, in particolare dei *Rerum vulgarium fragmenta*. Avviato a gennaio 2020, ITINERA vede collaborare mediolatini, filologi italiani, italianisti, paleografi, codicologi ed esperti di *Digital Humanities* e ha tra i suoi obiettivi la costruzione di una banca dati consultabile *online* in *open access*, in cui confluiranno le ricerche prosopografiche sui corrispondenti di Petrarca, quelle sui testi insieme ai dati codicologici e paleografici relativi alla tradizione manoscritta dei *Rerum vulgarium fragmenta*.

## 2. I CORPORA

Le ricerche di ITINERA insistono su tre *corpora* principali:

1. le corrispondenze di Petrarca, latine e volgari, in prosa e in versi (*Epystole; Familiarium rerum libri; Res seniles; Sine nomine; Disperse*; lettere a Petrarca; tenzoni e testi poetici genericamente di scambio, eventualmente inclusi nei *Rerum vulgarium fragmenta*);
2. i corrispondenti diretti di Petrarca;
3. la tradizione manoscritta dei *Rerum vulgarium fragmenta*, con speciale attenzione alla prima e più antica, entro il primo decennio del secolo XV.

Tenendo conto di esperienze e progetti analoghi e degli standard descrittivi in uso, sono state preliminarmente discusse la tipologia di dati da estrarre e le modalità di raccolta, organizzazione e interrogazione degli stessi. Un censimento preliminare e primi prototipi di descrizione hanno permesso quindi di individuare classi e categorie di descrizione sulla base della quali sono stati successivamente stabiliti i requisiti del *software* da adottare per lo sviluppo della banca dati.

## 3. IL DATABASE: TECNOLOGIA E PROGETTAZIONE

Per lo sviluppo della banca dati, ITINERA adotta un *framework* di *editing open source* sviluppato da Daniele Fusi, Cadmus, basato su una modellizzazione dinamica e modulare dei dati che consente di organizzarli in *set* indipendenti e componibili in strutture più complesse. Dal momento che ITINERA non descrive soltanto testi, ma raccoglie anche dati biografici, paleografico-codicologici ed altri tipi di dati, la struttura semantica di Cadmus risulta particolarmente adeguata alle sue esigenze.

Uno dei maggiori vantaggi di Cadmus risiede infatti nella possibilità di creare modelli logici senza vincoli di costruzione: non dipendendo da sintassi preordinate, i dati immessi in questi modelli possono successivamente essere esportati in formati diversi (API, TEI, CSV, TXT, PDF, ecc.), anche in RDF secondo ontologie standard. La progettazione in Cadmus è inoltre facilitata da un’interfaccia che permette di visualizzare la struttura dei dati già in fase di definizione dei modelli.

L'unità di base di Cadmus è la *part*, un *set* di dati autonomo e virtualmente indipendente strutturato secondo un preciso modello; ogni *part* contiene un certo numero di *properties* ed eventualmente *classes* (*bricks*); la somma di più *part* costituisce un *item*. Un *item* è dunque un 'contenitore' che assembla *part* funzionali alla rappresentazione di uno specifico oggetto.

Se il vantaggio di un *framework* improntato all'*open modeling* risiede nella possibilità di creare *item* altamente specifici e strutturati, il carattere modulare di Cadmus permette l'uso trasversale o il riuso di parti ed elementi già formalizzati (*composite modeling*): *properties* e *classes* largamente generalizzabili possono cioè essere sfruttati per la realizzazione di nuove *part* o *item*, evitando la difficoltà di progettare nuovi schemi di descrizione quando esistano soluzioni efficaci già formalizzate. Le caratteristiche dell'*open* e del *composite modeling* svincolano dunque la progettazione dall'uso di *set* predefiniti di modelli, non necessariamente adeguati alle esigenze della ricerca.

#### 4. IL DATABASE: I MODELLI

Il *data modeling* ha condotto alla formalizzazione di tre *item*, corrispondenti alle tre 'entità' che costituiscono i *corpora* sui quali insiste ITINERA:

- *Text* per i testi (*Metadata, Identifiers, Literary Work Info, Letter Info, Witnesses, Asserted Chronotopes, Related Text, Related Person, Events, Bibliography, Note*);
- *Person* per le persone (*Metadata, Identifiers, Names, Person Info, Person Works, Events, Bibliography, Note*);
- *Manuscript* per i manoscritti (*Metadata, Identifiers, Shelfmarks; Sheet Labels; Material Description, Watermarks, Layouts, Hands, Edits, Decorations, Bindings, Contents, Loci, Poem Ranges, History, Events, Bibliography, Note*).

La riusabilità di modelli e sotto-modelli ha permesso di adottare alcune parti in tutte e tre le tipologie di *item*, ma l'impiego trasversale di *bricks* è più frequente nella composizione di singole *part*, come nel caso delle informazioni cronotopiche (*Chronotopes*) o dell'indicazione di fonti a sostegno dei dati inseriti (*Document Reference*).

La modellizzazione di *Bibliography*, comune a tutti gli *item*, ha condotto all'elaborazione di un database bibliografico esterno e autonomo, in cui ora viene raccolta e archiviata la bibliografia usata per la creazione dei singoli *item*, che la richiamano al bisogno ricorrendo ad una *key*. Tale banca dati bibliografica ha fornito lo spunto per un *output* inizialmente non previsto dal progetto: la pubblicazione della prima bibliografia petrarchesca *online*.

#### 5. LOD MADE EASY

Se la costruzione di grafi LOD è ormai imprescindibile e non costituisce in sé una novità, il grande acquisto del progetto sta nell'aver rovesciato la prospettiva da cui, normalmente, quei grafi si costruiscono.

Elevando il grado di astrazione dei modelli descrittivi, Cadmus genera un grafo semantico **a posteriori**: i dati inseriti negli *item* progettati secondo i modelli sono mappati e proiettati automaticamente in grafi semantici ricorrendo ad ontologie standard (CIDOC-CRM). In questo modo il compilatore non ha bisogno di editare i dati 'nodo per nodo': tale prospettiva garantisce, per esempio, che le risorse digitali siano create in modo agevole e 'human friendly' anche da persone non necessariamente esperte di IT.

L'approccio di ITINERA è stato dunque creare, da un lato, modelli articolati e *human-friendly*, altamente componibili e riusabili; dall'altro, un sistema che automaticamente li selezionasse e mappasse per produrre da sé un grafo completo. Tale approccio, reso possibile da un modello virtuoso di collaborazione e dialogo tra esperti di DH e umanisti tradizionali, garantisce un mezzo straordinario per il raggiungimento degli obiettivi specifici del progetto, che punta a ricostruire una rete di relazioni su più livelli, a partire da una base di dati molto estesa e molto strutturata.

#### 6. WORKFLOW: METODI E STRUMENTI

Il piano di lavoro per la compilazione della banca dati è stato stabilito tenendo conto di obiettivi, tempi e risorse del progetto, attenendosi in particolare al metodo MoSCoW (*Must Have, Should Have, Could Have, Would Have*). La schedatura dei *corpora* è stata quindi distribuita tra le quattro unità di ricerca secondo le specifiche competenze di ciascuna. Ad essa collaborano, con ruoli diversi, membri del progetto e collaboratori esterni: ai ruoli assegnati a ciascuno corrispondono diversi livelli di autorizzazione nella banca dati. Preliminarmente formati attraverso appositi seminari (registrati e resi poi disponibili), i compilatori fanno riferimento a specifiche linee guida per il *data entry* predisposte in formato *wiki*, consultabili su *GitHub* (<https://petrarchsitinera.github.io/linee-guida/>) e disponibili in versione *offline*.

L'*authority control* è garantito dal ricorso a *repositories* pubblici appositamente allestiti, con cataloghi per i nomi di persona, i titoli dei testi e le segnature di manoscritti: tenendo conto delle nuove immissioni, le voci di autorità vengono aggiornate periodicamente dai responsabili individuati.

La revisione e il controllo di qualità, che avvengono mensilmente, prevedono che i dati inseriti siano verificati e validati da una persona diversa dal compilatore, e sono affidati, in particolare, ai responsabili di unità del progetto e ad altri membri *senior*.

Il *workflow* prevede che le quattro unità di ricerca e i collaboratori procedano alla compilazione degli *item* assegnati in modo parallelo e autonomo. Incontri di controllo trimestrali verificano lo stato di avanzamento dei lavori per individuare eventuali criticità e procedere alla loro risoluzione. Per garantire fluidità al *data entry*, inoltre, ITINERA ricorre a Click-up, un *software* di gestione dei progetti grazie al quale i compilatori possono segnalare dubbi, problemi o difficoltà incontrati durante il lavoro, indicando la tipologia della segnalazione e assegnandola ai responsabili deputati alla risoluzione di quello specifico problema.

## 7. PUBBLICAZIONE DEI DATI E OUTPUT

ITINERA pubblicherà la banca dati sul sito del progetto a *data entry* completato, comunque entro la conclusione del finanziamento (data prevista: gennaio 2024): al momento è stato coperto l'intero *corpus* dei testi, mentre sono in fase di compilazione gli *item* relativi a persone e manoscritti. Oltre a rendere interrogabili le schede descrittive che traducono i dati immessi nei singoli *item*, la *query* sviluppata permette l'incrocio e la parametrizzazione dei dati – raccolti secondo modelli che ne garantiscono la granularità anche con questo scopo – e prevede modalità di ricerca 'guidata', con domande preimpostate, pensate per favorire utenti poco esperti.

Sono in fase di sviluppo *output* che hanno l'obiettivo di valorizzare, anche graficamente, le possibilità di interrogazione e di impiego dei dati raccolti, in particolare mappe geografiche e *timelines* dinamiche, che saranno consultabili dal sito del progetto dove troverà collocazione anche la bibliografia petrarchesca.

## 8. DIGITAL E TRADITIONAL HUMANITIES

La progettazione dello strumento digitale ha coinciso con un dialogo continuo e serrato tra chi conosceva i *corpora* da descrivere e le modalità di ricerca che si intendevano offrire su di essi e chi quelle modalità doveva formalizzare a livello tecnico e tecnologico: un costante sforzo 'traduttivo', perciò, di conoscenze specifiche e specialistiche che conducessero a formalizzazioni adeguate agli standard e funzionali alla ricerca. Rovesciando la prospettiva dalla quale i membri dell'*équipe* guardavano inizialmente alla costruzione dello strumento – che coincideva in fondo con il punto di vista dell'utente-ricercatore, più abituato ad usare le banche dati che a progettarle – la collaborazione continuativa con i *digital humanist* ha comportato un acquisto metodologico fondamentale, determinando il passaggio da un *data modeling* condotto soprattutto avendo a mente la presentazione e la visualizzazione delle informazioni ad un altro incentrato piuttosto su strutture che rendessero i dati computabili.

## BIBLIOGRAFIA

- [1] Brovia, Romana, e Alessia Valenti. «La schedatura delle lettere di Petrarca nel progetto ITINERA». Bartoli, Elisabetta, Martina Pavoni, Cristiano Amendola, e Valeria Giovanna Nitti (eds.). *Le nuove frontiere del "Dictamen"*. Studi, edizioni in corso e riflessioni metodologiche sull'epistolografia medievale (sec. XII-XV). Firenze: SISMEL - Edizioni del Galluzzo, i.c.s.
- [2] Ciotti, Fabio. «La rappresentazione digitale del testo: il paradigma del markup e i suoi sviluppi». Perilli Lorenzo e Domenico Fiormonte (eds). *La macchina nel tempo. Studi di informatica umanistica in onore di Tito Orlandi*: 63-90. Firenze: Le Lettere, 2011.
- [3] De Nisco, Nicola, Sandra Gorla, e Alessia Valenti. «Un database per Petrarca e i suoi contemporanei: corrispondenze, corrispondenti e circolazione manoscritta». Manganaro Andrea, Giuseppe Traina, e Carmelo Tramontana (eds.). *Letteratura e potere/poteri. Atti del XXIV Congresso Nazionale dell'ADI, Catania, 23-25 settembre 2022*. Roma: Adi editore, 2023. <https://www.italianisti.it/pubblicazioni/atti-di-congresso/letteratura-e-potere>
- [4] De Nisco, Nicola, Sandra Gorla, e Alessia Valenti. «Petrarca tra lettere, corrispondenti e copisti: notizie dal progetto Itinera». *Linguistica e Letteratura XLVI/1-2 (2021)*: 107-143.
- [5] De Nisco, Nicola, Sandra Gorla, e Alessia Valenti. «Una banca dati per Petrarca e il suo tempo: criteri, modelli e obiettivi». *Digitalia 2 (2021)*: 67-90.
- [6] Fusi, Daniele. «Sailing for a Second Navigation: Paradigms in Producing Digital Content». *Seminari romani di cultura greca 7 (2018)*: 213-276.
- [7] Gnoli, Claudio e Carlo Scognamiglio. *Ontologia e organizzazione della conoscenza: introduzione ai fondamenti teorici dell'indicizzazione semantica*. Lecce: Pensa Multimedia, 2008.

- [8] Hyvönen, Eero. 2020. «Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery». *Semantic Web* 11 (1): 187–93.
- [9] Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham: Ashgate, 2015.
- [10] Van Deursen, Davy, Chris Poppe, Gäetan Martens, Erik Mannens e Rik Van de Walle. «XML to RDF conversion: A generic approach». *2008 International conference on automated solutions for cross media content and multi-channel distribution*: 138-144. Los Alamitos: IEEE, 2008.

# Per l'edizione digitale delle lettere familiari di Verga: un esperimento di analisi semantica

Giuseppe Zappalà<sup>1</sup>, Antonio Di Silvestro<sup>2</sup>, Pietro Sichera<sup>3</sup>, Daria Spampinato<sup>4</sup>

<sup>1</sup> Dipartimento di Scienze Umanistiche, Università degli Studi di Catania, Italia – giuseppe.zappala1@unict.it

<sup>2</sup> Istituto di Scienze e Tecnologie della Cognizione (CNR-ISTC), Catania, Italia – silvanto@unict.it

<sup>3</sup> Ministero dell'Istruzione e del Merito, Italia – pietro.sichera@posta.istruzione.it

<sup>4</sup> Istituto di Scienze e Tecnologie della Cognizione (CNR-ISTC), Catania, Italia – daria.spampinato@cnr.it

## ABSTRACT\*

Il presente lavoro riguarda l'edizione digitale delle lettere familiari di Giovanni Verga, realizzata a partire dalla codifica in XML/TEI delle stesse. Viene descritto lo stato del progetto, delineando le specifiche scelte adottate in fase di codifica, e viene proposto un esperimento di analisi condotto a partire dalla marcatura semantica di particolari aree di interesse contenute nelle lettere. La codifica dell'epistolario ha consentito di dar conto dell'evoluzione linguistica e semantica del corpus epistolare verghiano e di definire liste esterne, collegabili secondo le pratiche dei *Linked Open Data* (LOD) a vocabolari controllati sul web (*GeoNames*, *VIAF*) e a risorse esterne di lessicografia elettronica (*VoDIM*<sup>1</sup>). L'obiettivo è stato quello di evidenziare il vantaggio di un'analisi computazionale anche a livello ermeneutico, ponendo particolare rilievo a quegli elementi della codifica che rimandano ad aspetti semantici, come la definizione di <abstract> e <keywords>.

## PAROLE CHIAVE

Digital Scholarly Edition; lessico familiare; epistolario; analisi semantica; Verga

## 1. INTRODUZIONE

Sulla scorta dell'*Edizione Nazionale delle Opere di Giovanni Verga*<sup>2</sup> e in linea con i lavori avviati attraverso *DEMOTICON*[4], il progetto della codifica in XML/TEI delle lettere familiari di Giovanni Verga mira a fornire un modello di codifica digitale per l'intera opera epistolare dell'autore, servendosi dei metodi e degli strumenti delle edizioni scientifiche digitali sia in termini di analisi che di visualizzazione dei testi [7]. Ciò al fine di favorire una lettura in senso dinamico del testo, anche in relazione all'intero corpus dell'autore, e l'integrazione di materiale epistolare ancora inedito. Avvalendosi di precedenti progetti internazionali sull'edizione digitale degli epistolari<sup>3</sup>, la codifica qui proposta si è basata sull'edizione critica a stampa delle *Lettere alla famiglia* [9] di Giovanni Verga, edizione peculiare sia per l'ampio arco cronologico coperto che per le caratteristiche linguistiche e semantiche presenti in una modalità di scrittura meno controllata da parte dell'autore. Si tratta di un gruppo di 194 lettere, scritte fra il 1862<sup>4</sup> e il 1880, pubblicate secondo un criterio conservativo, con l'obiettivo di riprodurre gli aspetti più caratteristici della *facies* linguistica originaria della pagina d'autore [9: 13].

A partire dalla sezione *CorrespondenceDescription* del manuale TEI<sup>5</sup>, con l'integrazione dei moduli <correspDesc> e <correspContext>[8], l'analisi del contenuto ha riguardato 'luoghi', 'persone', 'organizzazioni', 'opere' e peculiari aspetti linguistici e semantici, con l'obiettivo di permettere l'ampliamento dell'edizione critica a stampa attraverso strumenti che possano fornire una più agevole e differente visualizzazione dei materiali, e che quindi siano in grado di collegare i diversi elementi dell'epistolario familiare con le altre opere dell'autore.

---

\* Il presente contributo è frutto di un costante confronto fra autori e autrici nelle fasi di ideazione, progettazione, stesura e revisione e della più completa condivisione di impostazioni e di contenuti: nondimeno, la responsabilità dell'introduzione e delle conclusioni va attribuita ad A. Di Silvestro e D. Spampinato, quella del paragrafo 2 a G. Zappalà, quella del paragrafo 3 a P. Sichera.

<sup>1</sup> *Vocabolario Dinamico dell'Italiano Moderno*. *Stazione lessicografica*, 2019, <https://www.stazionelessicografica.it/>

<sup>2</sup> *Edizione Nazionale delle Opere di Giovanni Verga*, <https://www.interlinea.com/interlinea/libri-collana-edizione-nazionale-delle-opere-di-giovanni-verga-53966.html>

<sup>3</sup> *Van Gogh Letters Project*, <https://vangoghletters.org/vg/>; *Bellini Digital Correspondence*, <https://bellinidigitalcorrespondence.cnr.it/>; *The Project WeGA*, <https://www.weber-gesamtausgabe.de/en/Index>

<sup>4</sup> La prima lettera del corpus è datata 1851, ma non è stata considerata in questa analisi perché isolata dalle altre.

<sup>5</sup> 2.4.6 *CorrespondenceDescription*. In *P5: Guidelines for Electronic Text Encoding and Interchange* (Version 4.5.0. Last updated on 25th October 2022, revision 3e98e619e), <https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD44CD>



## 2. UN ESPERIMENTO DI ANALISI SEMANTICA E LESSICALE

Inspirandosi agli obiettivi del progetto *VIVe*<sup>6</sup> per la costruzione di un dizionario del verismo, e collegandosi al noto interesse lessicografico di Verga (testimoniato dalla vasta corrispondenza con amici scrittori, *in primis* Capuana) [2], il presente contributo mira a definire un percorso di analisi semantica e lessicale all'interno dell'epistolario verghiano a partire dall'edizione codificata in XML/TEI attraverso l'editor *Oxygen*. Per far ciò, sono stati individuati elementi testuali che rimandano a liste esterne per 'luoghi', 'persone', 'opere' e 'termini' (Fig. 1), collegabili secondo le pratiche dei *Linked Open Data* (LOD) a vocabolari controllati sul web (*GeoNames*, *VIAF*) e a risorse esterne di lessicografia elettronica (*VoDIM*<sup>7</sup>). Questo al fine di prevedere, in fase di visualizzazione e ricerca, una lettura dinamica e multipla dell'opera verghiana sia in senso diacronico che sincronico e un arricchimento dei riferimenti a luoghi e tempi in cui visse e operò Verga, oltre che l'ampliamento degli elementi di conoscenza gravitanti attorno all'epistolario verghiano e il possibile collegamento con altre risorse e documenti dell'autore.

Nello specifico, per ciò che concerne gli aspetti lessicali, è stata prevista un'apposita lista esterna a cui rimandano i termini codificati nelle singole lettere attraverso il tag `<distinct>` e l'attributo di collegamento alla lista esterna `@ana`. La *ListTerm* si presenta attualmente suddivisa in due liste principali: la prima, in cui si codificano tecnicismi medici o giuridici identificati dal valore *tecnicism* dell'attributo `@ana` e dallo specifico termine di analisi *medterm* o *legalterm*; la seconda, in cui si codificano varianti diatopiche identificate dal valore *diatopicvariant* e da specifici termini di analisi per differenziare regionalismi e forestierismi (rispettivamente codificati come *regionalism* e *foreignism*).

Per quanto riguarda i tecnicismi, si fornisce una duplice informazione attraverso un rimando primario alla pagina della stazione lessicografica *VoDIM*, scegliendo come prima definizione quella riportata dal Tommaseo Bellini, e attraverso un rimando secondario all'eventuale nota esplicativa presente nell'edizione critica a stampa (Fig. 1).

```
<list ana="tecnicism">
  <interp xml:id="medterm">Lista di termini tecnici medici o con riferimento a pratiche di medicina alternativa.</interp>
  <interp xml:id="legalterm">Lista di termini giuridici</interp>

  <label ana="#legalterm" xml:id="mandamento">
    <note source="https://www.stazionelessicografica.it/ricerca?str=mandamento" n="TB.Mandamento.5"/>
    <term xml:lang="it">mandamento</term>
    <gloss>
      5. T. †† Ora abbiamo il Mandamento per Circostrizione amministrativa di territorio.
      E se la barbarie fosse unita, l'Italia sarebbe già bell'e fatta.
    </gloss>
  </label>
  <item>
    <note source="LAF-ListBibl.xml#LAF" resp="LAF-ListEditor.xml#ADS" n="V.5"/>
    <gloss xml:lang="it" target="#mandamento">
      «Circostrizione amministrativa intermedia tra il circondario e il Comune, in vigore in Italia fino al 1923».
      In diritto, «Circostrizione giudiziaria entro cui il pretore esplicava le proprie funzioni» (Zingarelli).
    </gloss>
  </item>
</list>
```

Figura 16. ListTerm: lista di tecnicismi nell'epistolario familiare di Verga, con rimandi a risorse lessicografiche esterne.

```
<list ana="diatopicvariant">
  <interp xml:id="regionalism">Lista di regionalismi o incertezze linguistiche derivanti dall'influenza regionale</interp>
  <interp xml:id="foreignism">Lista di forestierismi</interp>

  <label ana="#regionalism" xml:id="dietetropunto">
    <note source="https://books.google.it/books?id=jtFFAAACAAQ&printsec=frontcover&hl=it#v=onepage&q=retipuntu&f=false" n="TR.810.Retipuntu"/>
    <term xml:lang="it">retepuntato</term>
    <gloss>
      Cucitura in cui l'ago, in ogni punto successivo, si ripianta o a metà del punto precedente: punto addietro, o più vicino o giusto nel foro dov'è finito il punto precedente: impuntura rincrunata o punto cieco.
    </gloss>
  </label>
  <item>
    <note source="LAF-ListBibl.xml#LAF" resp="LAF-ListEditor.xml#ADS" n="XIV.4"/>
    <gloss xml:lang="it" target="#dietetropunto">
      Dal sic. retipuntu, e quindi 'cucito col dietropunto'
    </gloss>
  </item>
</list>
```

Figura 17. ListTerm: lista di varianti diatopiche nell'epistolario familiare di Verga, con rimandi a risorse lessicografiche esterne.

Per quanto riguarda le varianti diatopiche, si fornisce anche in questo caso una duplice informazione, attraverso un primo rimando a vocabolari siciliani dell'800 (Macaluso-Storaci, Traina) e un secondo rimando all'eventuale nota esplicativa presente nell'edizione critica a stampa (Fig. 2).

<sup>6</sup>*Vocabolario dell'Italiano Veristico*, [progettoviver.it/progetto.html](http://progettoviver.it/progetto.html)

<sup>7</sup>*Vocabolario Dinamico dell'Italiano Moderno*. Stazione lessicografica, 2019, <https://www.stazionelessicografica.it/>

Per gli aspetti semantici, sono invece state previste specifiche modalità di codifica all'interno della singola lettera, che rimandano a un *tagset* definito a livello di `<teiHeader>` (Fig. 3). Nel `<back>` della singola lettera (Fig. 4) è stato previsto l'inserimento dell'`<abstract>` per permettere una contestualizzazione e un riassunto del contenuto, a cui fa seguito un elenco di `<keywords>` all'interno del modulo `<textClass>`, codificate a partire dal *tagset* dichiarato nel `<teiHeader>`, per permettere una successiva indagine e visualizzazione dell'epistolario per temi. I due elementi codificati rappresentano un importante strumento di codifica e analisi dell'epistolario, in quanto permettono di operare a livello semantico nell'analisi e nello studio della singola lettera e dell'intero corpus, ripensando il percorso di lettura da parte dell'utente sia del singolo epistolario codificato che, in futuro, di altre opere dell'autore, e predisponendo in tal modo un utile mezzo per l'interoperabilità semantica fra dati testuali provenienti da lettere o documenti differenti con il rimando a una specifica tassonomia.

```
<classDecl>
  <taxonomy xml:id="semanticanalysis">
    <category xml:id="absence"><catDesc>Riferimenti alla distanza dall'ambiente familiare.</catDesc</category>
    <category xml:id="correspondence"><catDesc>Riferimenti ad altre corrispondenze.</catDesc</category>
    <category xml:id="daylife"><catDesc>Riferimenti ad attività quotidiane o lettere di saluti e informazioni generali.</catDesc</category>
    <category xml:id="editorial"><catDesc>Riferimenti ai rapporti editoriali e a specifici editori</catDesc</category>
    <category xml:id="family"><catDesc>Riferimenti alla vita del nucleo familiare.</catDesc</category>
    <category xml:id="familybusiness"><catDesc>Riferimenti agli affari economici del nucleo familiare.</catDesc</category>
    <category xml:id="health"><catDesc>Riferimenti alle condizioni di salute proprie o dei familiari.</catDesc</category>
    <category xml:id="journey"><catDesc>Riferimenti ai viaggi e agli spostamenti intrapresi dallo scrittore.</catDesc</category>
    <category xml:id="materialassets"><catDesc>Riferimenti al denaro, alle spese alle necessità economiche.</catDesc</category>
    <category xml:id="residence"><catDesc>Riferimenti ai cambi di residenza o domicilio da parte dello scrittore.</catDesc</category>
    <category xml:id="socialstruggle"><catDesc>Riferimenti al desiderio di affermazione sociale.</catDesc</category>
    <category xml:id="study"><catDesc>Riferimenti ai propri studi e ai materiali di studio.</catDesc</category>
    <category xml:id="works"><catDesc>Riferimenti alle proprie opere letterarie.</catDesc</category>
  </taxonomy>
</classDecl>
```

Figura 18. Tassonomia di *keywords* identificate ed esplicitate a livello di `<classDecl>` nel *TeiHeader*.

```
<abstract>
  <p>
    Verga informa la madre delle condizioni del nuovo appartamento in cui vive. Inoltre, fa riferimento alla commedia "Rose caduche", invitando Maro a non insistere con Niceforo e Salvini qualora non la reputassero adeguata al pubblico catanese. Si informa di alcuni affari legali di cui si sta occupando Maro e chiede del denaro per alcune spese.
  </p>
</abstract>
<textClass>
  <keywords>
    <term ana="familybusiness"></term>
    <term ana="materialassets"></term>
    <term ana="residence"></term>
    <term ana="editorial">Niceforo, Salvini</term>
    <term ana="works">Rose caduche</term>
  </keywords>
</textClass>
```

Figura 19. `<abstract>` e `<keywords>` nella lettera identificata come LAF27.

A partire dal saggio introduttivo all'edizione [9: 15-42], si è infine scelto di indagare a livello semantico il contenuto delle singole lettere per le categorie relative al desiderio di affermazione sociale (*socialstruggle*) e alle necessità economiche a cui fa riferimento lo scrittore (*materialassets*), caratterizzanti in senso quantitativo e qualitativo l'epistolario. Il tema cardine dell'intera silloge risulta, infatti, essere quella 'lotta per l'esistenza' che coerentemente emergerà nel nucleo familiare dei *Malavoglia*. Un tema che si traduce a livello epistolare nell'uso insistito di uno specifico vocabolario che rimanda a più riprese all'area semantica della lotta, nella necessità di «riuscire a qualche cosa» [9: 127, lettera XXVI del 12 giugno 1869, alla madre Caterina Di Mauro Barbagallo]. Le lettere alla madre e ai fratelli presentano ripetutamente il tema, da una parte come giustificazione che l'autore trova nei confronti dei familiari per i «sacrifici che costo» [9: 93, lettera XVII del 7 Maggio 1869, al fratello Mario Verga], dall'altra in quanto momento di confessione intima delle difficoltà nell'integrarsi nella società del tempo con l'obiettivo di «vivere in mezzo a questo movimento incessante, farsi conoscere e conoscere, respirarne l'aria insomma» [9: 93, lettera XVII del 7 Maggio 1869, al fratello Mario Verga].

Si tratta di tematiche strettamente connesse fra loro, che trovano però una propria specificità nel 'romanzo di formazione' verghiano, riguardando a volte un'area prevalentemente privata e altre un'area pubblica. La lotta fra il bene e il male, la lotta per la sopravvivenza, i sacrifici per accedere a una società da cui si sente irrimediabilmente attratto, ma che non sente come propria, occupano una parte considerevole delle lettere alla madre nel momento in cui avviene il trasferimento a Firenze e Verga tenta di inserirsi nella società letteraria del tempo, nei circoli della cultura. Il denaro e gli affari coprono, invece, un'area soprattutto privata, in quanto fanno riferimento agli affari familiari, a questioni legali, alla ricerca di una propria autonomia economica. È chiaro, altresì, che le due aree sono strettamente connesse fra di loro nell'ambito dell'inserimento di Verga in una società nuova e altra rispetto a quella da cui proviene, una lotta interiore ed esteriore per

accedere al mondo dei ‘vincitori’. L’insistenza sul denaro è dunque accostabile non soltanto alle necessità pratiche di Verga fuori dalla sua Catania, ma anche alla necessità di inserirsi in un ambiente che non gli appartiene, ma che egli sente come proprio, un ambiente fatto di amici e nemici, in cui lottare per emergere e sopravvivere.

```

Io sto benissimo; soltanto <seg type="semanticanalysis" ana="socialstruggle"> queste conoscenze </seg> mi hanno
imposto delle <seg type="semanticanalysis" ana="materialassets">spese</seg> indispensabili di guanti e colli di
camicia. Ecco perché il mio <seg type="semanticanalysis" ana="materialassets">bilancio</seg> si trovò
sconcertato in modo che dovetti scrivervi di mandarmi <seg type="semanticanalysis" ana="materialassets">qualche
cosa</seg> alla prima del mese.
Del resto vi posso assicurare che faccio accuratamente i miei <seg type="semanticanalysis" ana="materialassets">
conti</seg> e non spendo un <seg type="semanticanalysis" ana="materialassets">soldo</seg> a capriccio.

Vorrei che <persName ref="LAF-ListPerson.xml#MarioV"><addName type="nickname">Maro</addName></persName> dicesse
a <persName ref="LAF-ListPerson.xml#Niceforo">Niceforo</persName> se potrei mandargli una mia <rs type="comedy"
ana="LAF-ListOpere.xml#Rosecaduche"> commedia</rs> e se <persName ref="LAF-ListPerson.xml#Salvini">Salvini</
persName> potrebbe incaricarsi di rappresentarla durante il corso delle rappresentazioni che darà a <placeName
ref="LAF-ListPlace.xml#CT">Catania</placeName>. Ne attendo risposta al più presto.

```

Figura 20. Categorie semantiche individuate nell’epistolario: *materialassets* e *socialstruggle* nella lettera identificata come LAF23.

### 3. ANALISI QUANTITATIVA DEI DATI SEMANTICI

L’obiettivo della codifica semantica è stato dunque quello di analizzare la presenza e la distribuzione delle categorie semantiche individuate all’interno dell’epistolario familiare, dimostrando il vantaggio anche in termini ermeneutici di un’analisi computazionale. L’analisi è stata effettuata utilizzando il tool *NormaTEI*, software ideato all’interno del progetto *Bellini Digital Correspondence* e sviluppato attraverso la piattaforma di sviluppo 4D, che consente un controllo qualitativo e un’analisi quantitativa della codifica XML-TEI di più files, permettendo inoltre l’armonizzazione del *corpus*. Diventa così possibile fornire una prima conferma di quanto espresso da Savoca nell’introduzione all’edizione critica, riuscendo inoltre a dimostrare l’andamento regressivo dei temi analizzati nei tre periodi in cui sono state raggruppate le lettere (1862-1869; 1874-1875; 1877-1880). Pur coprendo un arco cronologico così ampio, la distribuzione quantitativa delle lettere risulta disomogenea, in quanto al primo periodo corrispondono 53 lettere, al secondo 106 e al terzo 28. Per tale ragione, si è scelto di procedere all’analisi quantitativa dei valori *materialassets* e *socialstruggle* – codificati all’interno del corpo della lettera nei tag <measure> e <seg> – per singolo periodo, registrando la presenza di una singola occorrenza del valore per lettera e rapportandola al totale delle lettere per periodo. Il valore emerso per *materialassets* è stato dell’85% nel primo periodo, del 59% nel secondo periodo e del 36% nell’ultimo periodo (Fig. 6), mentre per *socialstruggle* si registra una tendenza opposta, per cui la presenza di elementi codificati come appartenenti a tale categoria semantica risulta essere maggiore nell’ultimo gruppo di lettere, con un valore del 32% (Fig. 7).

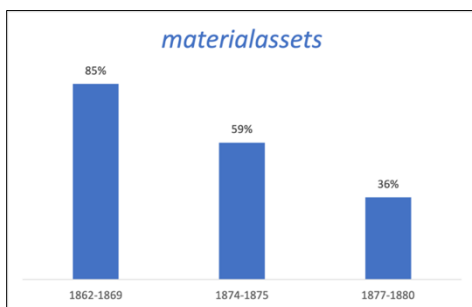


Figura 21. Percentuale *materialassets* per periodo.

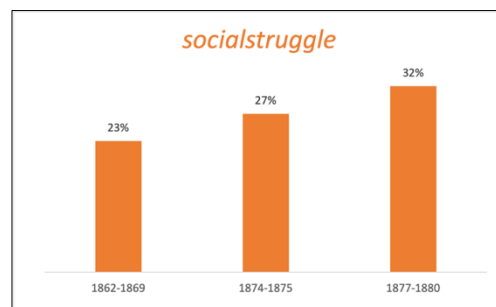


Figura 22. Percentuale *socialstruggle* per periodo.

Tale andamento opposto potrebbe spiegarsi con il progressivo miglioramento delle condizioni economiche di Verga grazie alla pubblicazione e al successo di alcune opere, a cui segue una maggiore preoccupazione rispetto alla propria affermazione sociale e all’opinione della società letteraria del tempo. L’analisi incrociata delle due categorie semantiche permette, inoltre, di affermare che il 67% delle lettere presenta come argomento *materialassets* e il 27% ha come argomento *socialstruggle*, per un complessivo 90% delle lettere che fanno riferimento alle due categorie e che dunque rimandano a quella lotta per riuscire in qualcosa di cui si è già parlato (cfr. *supra*).

Un’ulteriore analisi è stata condotta all’interno della categoria semantica del *materialassets* (Fig. 8), categoria quantitativamente più significativa all’interno del corpus, con una presenza dell’argomento in quasi il 70% delle lettere. Si tratta di una categoria semantica all’interno della quale sono stati codificati lemmi e porzioni testuali differenti, e che quindi raggruppa varianti (anche grafiche), sinonimi ed espressioni che rimandano alla medesima categoria del bisogno materiale

sia in senso personale sia in riferimento al gruppo familiare più ristretto. Si è scelto, così, di dar conto dei termini maggiormente rappresentativi della categoria semantica del *materialassets*, ovvero ‘lira’, ‘denaro’, ‘franco’, ‘spesa’, ‘affare’, ‘soldo’, di cui sono state analizzate e raggruppate le diverse forme e varianti grafiche. Ciò è stato possibile grazie alla normalizzazione delle forme linguistiche per le due categorie semantiche prese in considerazione, riuscendo così ad analizzarne la vitalità e la variabilità nell’arco dell’intero epistolario. Emerge così l’insistenza sul denaro in senso materiale, con una somma delle diverse forme di ‘lira’, ‘denaro’ (e ‘danaro’) e ‘franco’ che supera il 50%, mentre si registra un’interessante percentuale di occorrenze di ‘spesa’ e ‘affare’, che potrebbero rimandare ai diversi affari familiari in cui è impegnata la famiglia Verga e di cui lo scrittore chiede spesso aggiornamenti.

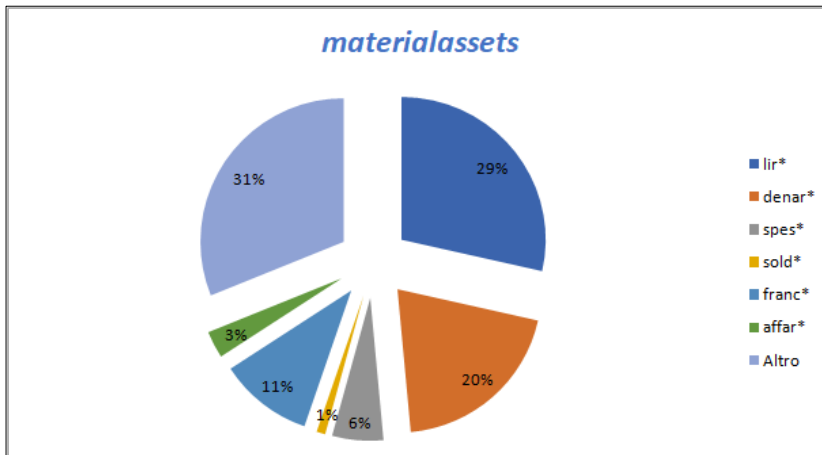


Figura 23. Forme codificate nella categoria semantica *materialassets*.

#### 4. CONCLUSIONI E SVILUPPI FUTURI

L’analisi distribuzionale condotta consente altri interessanti rilievi, allorché venga condotta un’analisi incrociata tra le *keywords* e i quattro destinatari principali delle lettere. A emergere è sempre una sostanziale prevalenza di *materialassets* e *socialstruggle*, che si concentrano nella quasi totalità nelle 108 lettere inviate alla madre e nelle 45 spedite al fratello Mario, vero e proprio *pater familias* alla morte di Caterina Di Mauro. Sulla funzione polarizzatrice dell’universo materno rispetto ai temi e agli aspetti portanti del vissuto dell’uomo e dello scrittore non è il caso di insistere ulteriormente. A incuriosire è invece soprattutto il fatto che, scrivendo alla madre, donna nient’affatto di studi avanzati, Verga parlasse copiosamente di questioni culturali ‘alte’ come quelle editoriali (*editorial* appare in 35 lettere), che avrebbero in realtà dovuto trovare spazio nelle missive allo zio Salvatore e soprattutto in quelle a Mario (unico dei fratelli ad aver conseguito una laurea), estremamente attento alle trattative per la pubblicazione delle opere di Giovanni, e coinvolto in alcuni casi nella ricopiatura dei racconti per l’invio agli editori (ad es. per *Iricordi del Capitano d’Arce*). Questa ‘anomalia’ si potrebbe motivare con un duplice ordine di considerazioni: 1) quando scrive alla madre, Verga si rivolge spesso a quello che egli chiama il “consiglio di famiglia” [cfr. 9: 335, lettera CXI del 19 maggio 1874, alla madre Caterina Di Mauro Barbagallo], comprensivo quindi di fratelli e sorelle; 2) le lettere allo zio Salvatore, altra figura colta della famiglia, sono quantitativamente esigue, e raramente incrociano periodi di fervida attività scrittorica ed editoriale.

In una prospettiva ermeneutica più generale, un’analisi semantica per *keywords* permette di valorizzare la natura complessa e stratificata, plurisensa e mai meramente informativa o referenziale, della scrittura epistolare. I livelli di senso che nella lettera si intersecano ne fanno centro di irradiazione di una pratica scrittorica che, in un autore come Verga, istituisce un contatto ‘diretto’ tra la comunicazione privata e quella letteraria, superando la distinzione tra il genere lettera e i generi narrativi (romanzo e novella) all’insegna di una universalizzazione dei temi che caratterizzano il vissuto dello scrittore e quello dei suoi personaggi.

Progetti internazionali di edizioni digitali degli epistolari permettono oggi di formulare l’auspicio di un’edizione nazionale digitale dell’epistolario verghiano [1,3,9]. In tal senso, si sta lavorando alla progettazione di un’interfaccia dell’edizione digitale delle lettere, con l’obiettivo di definire e costruire una risorsa digitale per lo studio della vita e dell’opera di Verga, collegabile a esperienze parallele di analisi e codifica dell’opera dell’autore [5]. La codifica semantica proposta sarà, inoltre, considerata come punto di partenza per un’analisi incrociata sulle medesime lettere o su altre categorie semantiche all’interno del corpus verghiano, e aperta da una parte a un’indagine su più livelli, che possa collegare l’intero corpus verghiano e risorse LOD, dall’altra a successive integrazioni di materiale ancora inedito e alla ridefinizione in tal senso del flusso autoriale e della memoria linguistica e letteraria verghiana.

## 5. RINGRAZIAMENTI

Si ringraziano Antonella Zammataro e Consuelo Borzi per il contributo nella codifica delle lettere, risultato dell'attività di tirocinio presso il CNR con la collaborazione del CINUM (Centro di Informatica Umanistica), nell'ambito del Corso di laurea in *Scienze del testo per le professioni digitali* dell'Università di Catania. Si ringrazia Christian D'Agata per il supporto e la revisione nella costruzione del modello di codifica proposto.

## BIBLIOGRAFIA

- [1] Alfieri, G. «Presentazione». *Annali della Fondazione Verga VI* (2013): 7-22.
- [2] Alfieri, G. «Verga e “il valore d’uso” nella lingua e nel dialetto, tra vocabolari siciliani, toscani e... non solo». *Bollettino del Centro Studi Filologici e Linguistici Siciliani*, 32 (2021): 117-146.
- [3] Branciforti, F. «Lo scrittoio del verista». In *I tempi e le opere di Giovanni Verga. Contributi per l'Edizione Nazionale*, a cura di F. Branciforti e G. Galasso. Firenze: Le Monnier, 1986.
- [4] Cristofaro, S., D'Agata, C., Di Silvestro, A., Palazzolo, G., Sichera, P., Spampinato, D. «Demoticon. Per un'edizione semantica dei Malavoglia». In *AIUCD 2021 - DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale. Raccolta degli abstract estesi della 10° conferenza nazionale*, a cura di F. Boschetti, A.M. Del Grosso, E. Salvatori: 471-473. Pisa: AIUCD, 2021.
- [5] Di Silvestro, A., D'Agata C., Canzoneri, G., Grasso, M. «Per un'edizione digitale delle opere di Giovanni Verga: rappresentazione del testo e modelli di codifica (con alcuni esempi sulle novelle)». *Rivista di Letteratura Italiana* 3 (2022): 45-66.
- [6] Pellino, S., Sichera P., Del Grosso, A.M., Spampinato, D. «Dalla codifica alla fruizione: l'edizione digitale Bellini Digital Correspondence». In *AIUCD 2022 – Proceedings*, a cura di F. Ciraci, G. Miglietta, C. Gatto (Eds.): 163-168. Bologna: Quaderni di Umanistica Digitale, 2022.
- [7] Pierazzo, E. *Digital Scholarly Editing: Theories, Models and Methods*. London: Routledge. [2015] 2016.
- [8] Stadler, P., Illutschko, M., Seifert, S. «Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>». *Journal of the Text Encoding Initiative* 9, (2019). <http://journals.openedition.org/jtei/1433>
- [9] Verga, G. *Lettere alla famiglia (1851-1880)*, a cura di G. Savoca e A. Di Silvestro. Acireale-Roma: Bonanno, 2011.



# Ease the collaboration making Scholarly Editions: the GreekSchools case study

Simone Zenzaro<sup>1</sup>, Angelo Mario Del Grosso<sup>2</sup>, Federico Boschetti<sup>3</sup>, Graziano Ranocchia<sup>4</sup>

<sup>1</sup> Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILS), Pisa, Italy – simone.zenzaro@ilc.cnr

<sup>2</sup> Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILS), Pisa, Italy – angelo.delgrosso@ilc.cnr

<sup>3</sup> Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILS), Pisa, Italy – federico.boschetti@ilc.cnr.it

<sup>4</sup> Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Italy – graziano.ranocchia@unipi.it

## ABSTRACT

CophiEditor is a Digital Scholarly Editing Web platform based on Domain Specific Languages (DSL-based DSE). We are developing this platform in the context of the GreekSchools-885222 ERC project in which it is being used for the *constitutio textus* of Philodemus of Gadara's Arrangement of the Philosophers. The digital papyrological edition of these texts is challenging in many ways, as most of the Herculaneum papyri are highly fragmentary because they are carbonized and damaged.

One of the main goals of the CophiEditor is to provide a full-fledged collaborative environment in order to support the scholar's editorial work. In this poster we show how the progress of CophiEditor eases the collaboration between scholars and the role played by the Web Annotation Data Model (WADM) for data representation and interchange.

## KEYWORDS

Domain Specific Languages; Computational Philology; Digital Philology; Web Annotation Data Model; DSE tools

## 1. THE COPHIEDITOR ANNOTATION SYSTEM

Making a Digital Scholarly Edition (DSE) is complex [1,2] and collaboration is difficult [3,4,5]. Our DSL-based DSE methodology to digital scholarly editing expects that tools for DSE should provide a familiar (or at least easy to learn) environment that facilitates the editing process.

This is one of the pillar ideas on which CophiEditor is grounded: enabling seamless collaboration in a Digital Scholarly Editing Web platform leveraging the advantages of Domain Specific Languages (DSL) [6].

The CophiEditor web platform is being developed in the context of the GreekSchools-885222 ERC AdG project<sup>1</sup> whose main goal is the production of an Open Access digital critical edition of Philodemus of Gadara's Arrangement of the Philosophers. In particular, the philologists in the project team are using our (still in development) web platform for the *constitutio textus* (e.i. textual editing) of the Herculaneum papyri towards a diplomatic and a literary transcription of the surviving fragments, together with a paleographic apparatus, a philological apparatus and an Italian translation.

The philologists are reconstructing the Greek text conveyed by the Herculaneum papyri, most of which are carbonized and highly damaged. To produce the edition, scholars are used to working together in the form of in-person workshops where different experts (papyrologists, scholars of ancient Greek, historians) meet and collaborate, focusing their intensive editorial work on specific parts of the text of the future publication. In this environment, a full-fledged collaborative and digital environment might accelerate the production of the scholarly edition.

CophiEditor, inspired by the Euphoria model [7], fits into the set of tools for DSEs that support collaboration, and the ability to annotate the text such as CATMA<sup>2</sup>, INCEPTION<sup>3</sup>, GATE workbench<sup>4</sup>, BRAT, Hypothes.is<sup>5</sup>.

CophiEditor is designed to both support collaboration (i.e. many participants for a single task) and cooperation (i.e. many participants for many sub-tasks). The aim of this contribution is to show how CophiEditor (among the other features) implements collaboration and the consequences for the scholars' work in the context of the GreekSchools project.

Since CophiEditor is a web application, no particular physical location is required to work on editing, commenting and annotating the text, but an electronic device able to connect to the Internet, making remote collaboration possible. Within the platform, scholars have access to the DSE working drafts and can edit or annotate each one by adding a comment to

---

<sup>1</sup> <https://greeschools.eu/>

<sup>2</sup> <https://catma.de/>

<sup>3</sup> <https://inception-project.github.io>

<sup>4</sup> <https://gate.ac.uk/>

<sup>5</sup> <https://web.hypothes.is/>

the text. The selection of a target text for a comment supports single and multiple selections, in case the comment extends over separate lines of text. The comments are visible to the other scholars who can reply to them and start a discussion thread, or add other comments to the editing text. Each comment update is propagated to all users connected to the platform who can henceforth participate in the discussion in real time. Since text is considered itself a particular type of annotation, changes to the edited text are also updated to other users in real time. The annotation model is based on the recursive data representation described in Omega [8], which allows comments to have maximum freedom regarding the granularity of the annotated text and, in particular, infinite depth regarding the hierarchy of annotations. This new feature of CophiEditor is implemented, from a technical point of view, in a microservice architecture in which the user interface is able to receive notifications from a service that collects all events of interest to the application and forwards them to browser clients via WebSockets that allow two-way communication between the service and the client<sup>6</sup>.

The annotation model implemented in CophiEditor is also compatible with the Web Annotation Data Model<sup>7</sup> (WADM) by design and makes it possible to derive the WADM representation of the comments. WADM is a well-known W3C standard that “describes a structured model and format to enable annotations to be shared and reused across different hardware and software platforms” and in our viewpoint, ensuring the compatibility with this standard is an important step toward interoperability and social [9] and assertive scholarly editing [10]. Moreover, since the texts of the edition are constantly changing until they are considered *ne varietur* (i.e. it must not be changed) by the editor, the comments may refer to some content that has been modified or that is not present anymore. In this case, the comment itself may be (or may be not!) considered obsolete, incorrect, or no longer relevant. WADM defines the concept of *state* that “provides the information needed to retrieve the correct representation” of the annotation. This way CophiEditor can manage comments to the text that have been changed to notify the edition editor.

The collaboration features described in this contribution are already implemented in the CophiEditor platform but, to achieve a full-fledged collaborative editing environment, the next steps we are going to take are: (1) the integration of Operational Transformation (OT) [11] for the edition texts, (2) a notification system. The OT approach will add the possibility to make concurrent changes to the same text by different users. The notification system will allow the users to be aware of comments threads or other even when they are working on other sections of the edition.

## 2. ACKNOWLEDGMENTS

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).

## REFERENCES

- [1] Fischer, F. 2013. “All Texts Are Equal, But... Textual Plurality and the Critical Text in Digital Scholarly Editions.” *Variants* 10: 77–92.
- [2] Bozzi, A., M.S.C. Bozzi, and G. Ferrari. 2019. *Percorsi Di Linguistica e Di Filologia Computazionali*. Edizioni ETS.
- [3] Bordalejo, B., and P. Robinson. 2015. “A New System for Collaborative Online Creation of Scholarly Editions in Digital Form.” In *1st Dixit Convension on Technology, Software, Standards for the Digital Scholarly Edition Workshop*. The Hague.
- [4] Robinson, P.. 2017. “Some Principles for Making Collaborative Scholarly Editions in Digital Form.” *DIGITAL HUMANITIES QUARTERLY* 11 (2).
- [5] Robinson, P. 2016. *The Digital Revolution in Scholarly Editing*. In: Crostini, B et al (eds.), *Ars Edendi Lecture Series*, vol. IV. Stockholm: Stockholm University Press.
- [6] Zenzaro, S., A. M. Del Grosso, F. Boschetti, and G. Ranocchia. 2022. “Verso La Definizione Di Criteri per Valutare Soluzioni Di Scholarly Editing Digitale: Il Caso d’uso GreekSchools.” In *AIUCD 2022*.
- [7] Boschetti, F., and A. M. Del Grosso. 2020. “L’annotazione di testi storico-letterari al tempo dei social media.” *Italica Wratislaviensia* 11 (1): 65–99.
- [8] Del Grosso, A. M., E. Giovannetti, and S. Marchi. 2017. “The Importance of Being... Object-Oriented: Old Means for New Perspectives in Digital Textual Scholarship.” In *Advances in Digital Scholarly Editing*, 1:269–74. Leiden: Sidestone Press.

---

<sup>6</sup> Usually the server can only send information to the client on demand, this approach makes it possible to send data to the clients on events that are not generated on the client.

<sup>7</sup> <https://www.w3.org/TR/annotation-model/>, <https://www.w3.org/TR/annotation-protocol/>



- [9] Siemens, R., M. Timney, C. Leitch, C. Koolen, A. Garnett. 2012. "Toward Modeling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media." *Literary and Linguistic Computing* 27 (4): 445–61.
- [10] Vogeler, G. 2019. "The 'Assertive Edition.'" *International Journal of Digital Humanities* 1 (2): 309–22..
- [11] Ellis, C. A., and S. J. Gibbs. "Concurrency control in groupware systems." *Proceedings of the 1989 ACM SIGMOD international conference on Management of data*. 1989.

## **2. Analisi computazionale dei testi**

# Zoo-KG: A knowledge graph of zoological data

Arnaud Barbe<sup>1</sup>, Molka Tounsi Dhouib<sup>2</sup>, Catherine Faron<sup>3</sup>, Marco Corneli<sup>4</sup>, Arnaud Zucker<sup>5</sup>

<sup>1</sup> Université côte d'Azur, CEPAM, CNRS, France – arnaud.barbe@univ-cotedazur.fr

<sup>2</sup> Université côte d'Azur, INRIA, CNRS, I3S France – dhouib@i3s.unice.fr

<sup>3</sup> Université côte d'Azur, INRIA, CNRS, I3S, France – faron@i3s.unice.fr

<sup>4</sup> Université côte d'Azur, CEPAM, CNRS, France – marco.corneli@univ-cotedazur.fr

<sup>5</sup> Université côte d'Azur, CEPAM, CNRS, France – arnaud.zucker@univ-cotedazur.fr

## ABSTRACT

In this paper, we present the ZooKG-Pliny knowledge Graph constructed from a manual annotation of Pliny's *Naturalis Historia* using concepts gathered in the thesaurus TheZoo. ZooKG-Pliny is based on a semantic model that formalizes knowledge about the annotations of zoological information in texts. ZooKG-Pliny allows the integration and the interrogation of relevant knowledge in order to support epistemologists, historians and philologists in their analysis of these texts and knowledge transmission through them.

## KEYWORDS

History of Zoology; Semantic annotation of Latin texts; RDF knowledge graph; Ancient animals; Linked data and vocabularies

## 1. INTRODUCTION

Historians and philologists have to deal every day with an enormous amount of textual resources. Despite the digitization efforts, the proposed tools do not meet the epistemological requirements. They are often limited to lexical and quantitative search on the data. Researchers express a real need for more intelligent tools in order to carry out more elaborate searches that require richer semantic annotation. A typical approach to integrating and exploiting text annotations is to formalize them as a knowledge graph based on linked data and Semantic Web models[1,2,3]. In this context, the objective of our work is to transform into a knowledge graph (ZooKG-Pliny) the manual annotations of *Naturalis Historia* by Pliny the Elder, made by only one linguist researcher and using concepts from TheZoo thesaurus. The ZooKG-Pliny allows the integration and querying of relevant knowledge to provide researchers with richer automated search capabilities to better meet their needs.

## 2. ZOO-KG: ZOOLOGY KNOWLEDGE GRAPH - PLINY'S NATURALIS HISTORIA

In this work, we considered the books 8 to 11 of Pliny's « *Naturalis Historia* » on ancient zoology. These four books are divided into 911 paragraphs, which are manually annotated by linguists with concepts from the TheZoo thesaurus. In order to automatically exploit these annotations, we lifted the text and the manual annotations into an RDF dataset and exploited it with SPARQL queries. The first step is to automatically generate an RDF dataset describing the semantics of the decomposition of Pliny's work into chapters and paragraphs. The second step consists in integrating the annotations of the text into this RDF dataset. Figure 1 illustrates an example RDF graph representing the fourteen paragraph of chapter 11 *Naturalis Historia*. Thanks to the link between the annotation and the TheZoo thesaurus, we can enrich our knowledge graph by exploiting other existing links in TheZoo. For example, the concept `idc:5066` which denotes a « Tiger » is a sub-concept of the one denoting « Pantherinae ».

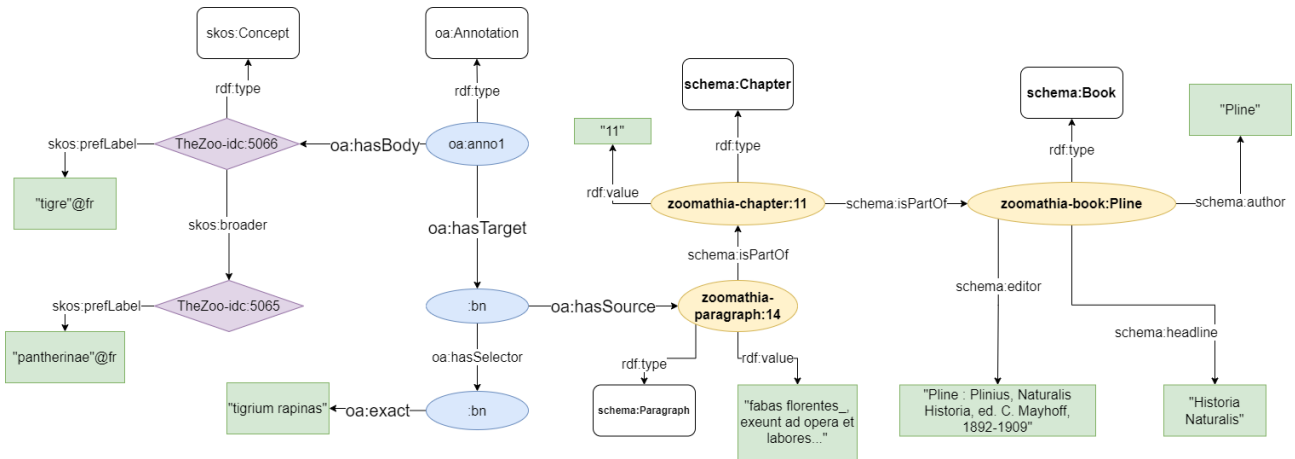


Figure 1. Example RDF graph representing the annotation of « igrium rapinas » in paragraph 14 of chapter 11 of Pliny's « Naturalis Historia ».

At the end of this process, we were able to automatically extract 11073 concepts from the expert's manual annotations, and generated 88184 RDF triples. Table 1 summarizes the number of manual annotations and other key statistics on the resulting knowledge graph.

Number of paragraphs	911
Number of comments	7283
Number of concepts to look for	13241
Number of concepts found	11073
Number of concepts not found	2478
Number of RDF triples generated	88184

Table 1. Number of entities per class in our knowledge graph.

To illustrate how we can exploit ZooKG-Pliny, we developed a set of SPARQL queries available on the project's Github repository. We first collected and discussed a set of competency questions (CQs) formulated by an expert in order to precisely understand their expectations and needs. As an example, Figure 2 presents researchers need to identify in the scientific literature animals that are capable of building a favorable habitat suitable for their needs.

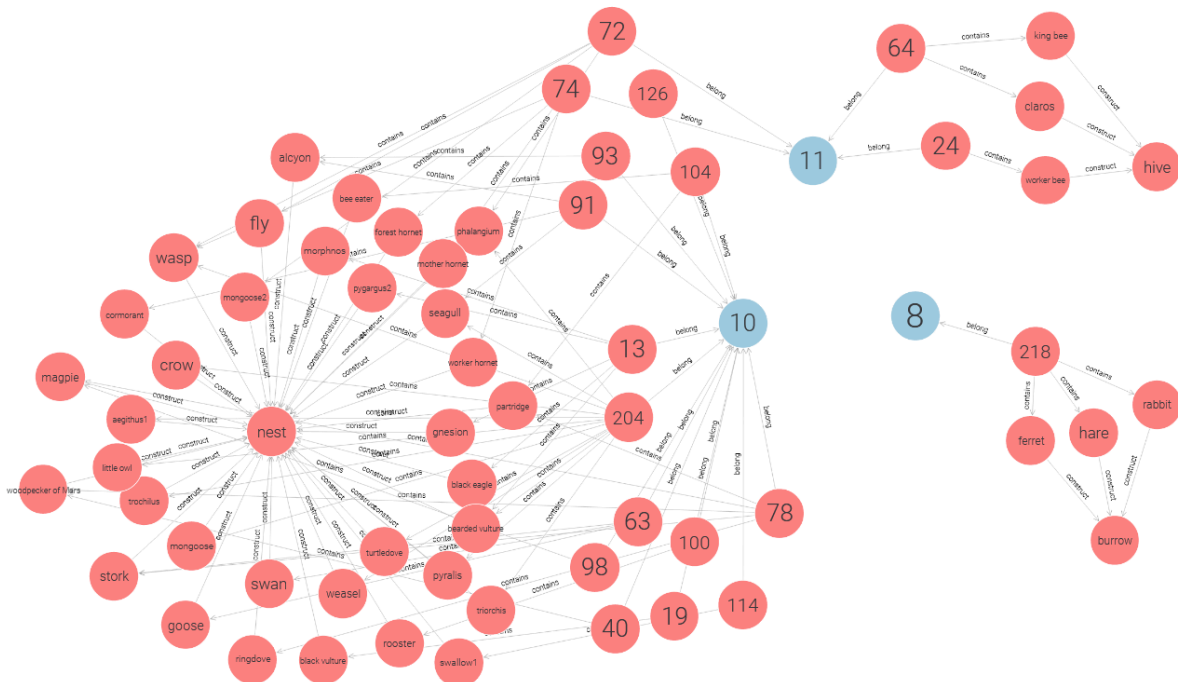


Figure 2. Example of visualization showing the set of animals and their associated constructions cited in Pliny's chapters.

### 3. CONCLUSION

ZooKG-Pliny is a knowledge graph formalizing manual annotations of four books (8 to 11) of the *Naturalis Historia* by Pliny the Elder. We released a fully automatic pipeline that structures manually extracted information from text and adds a semantic layer for smarter search. As perspectives of this work, we plan to automate or support the manual annotation task that remains tedious for experts and to develop an application that will make it easier for domain experts who are not semantic web specialists to use the constructed knowledge graph via more intelligible or intuitive visualization interfaces.

### REFERENCES

- [1] Hyvönen, Eero. "Digital humanities on the Semantic Web: Sampo model and portal series." *Semantic Web Preprint* (2022): 1-16.
- [2] Leyra, Irene Pajón, Arnaud Zucker, and Catherine Faron Zucker. "TheZoo: un thesaurus de zoologie ancienne et médiévale pour l'annotation de sources de données hétérogènes." *Archivum Latinitatis Medii Aevi* 73 (2015): 321-342.
- [3] Tounsi, Molka, Catherine Faron Zucker, Arnaud Zucker, Serena Villata, and Elena Cabrio. "Studying the history of pre-modern zoology with linked data and vocabularies." In *The First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*. 2015.

# *Do AIs dream of electric comics?*

## Generative AI models, digital memory, and creativity

Giorgio Busi Rizzi<sup>1</sup>

<sup>1</sup> Ghent University, Belgium – giorgio.busirizzi@ugent.be

### ABSTRACT

Generative AIs are models that, having been trained on large datasets of images, can create original art. Although we are still far from a hegemony of comics made this way, 2022 has already seen the emergence of a new generation of AIs, and with it an initial, yet already significant proliferation of AI-created comics and graphic novels (the most notable cases being *Fastwalkers*, Manouach 2022; *The Abolition of Man*, Grubaugh 2022; *Prompt. Conversations with AI*, McKean 2022; *Lovebot*, VV.AA., 2022).

This article examines the emergence of AIs generating images in the field of artistic creation. These models have the potential to automate certain aspects of the creative process, but also raise several ethical issues related to creativity, labor, and intellectual property. Through an overview of contemporary AIs and a survey of recent AI-generated comics, the article explores the impact of this new technology, discussing whether such creations can be considered “creative” and how they relate to existing artistic practices. Ultimately, the article aims to help navigate this new frontier of artistic creation.

### KEYWORDS

Comics; AI; generative AI models; creativity; memory

## 1. INTRODUCTION: UNPACKING ARTIFICIAL INTELLIGENCE

Artificial intelligence is a vast umbrella term that describes a range of software capable of performing actions related to perceiving, learning, inferring, and synthesizing information. AIs can encompass very different systems, first and foremost in terms of the complexity of their learning process: although machine learning and deep learning are often equated in the general discourse, the latter is a subset of the former, meant to handle larger and more complex datasets by using artificial neural networks with multiple layers (hence the term “deep”). These networks are inspired by the structure and functioning of the human brain and consist of interconnected nodes, or artificial neurons, that process and transmit information. In both cases, their learning process can be supervised, unsupervised, and based on reinforcement. In the first case, the algorithm is trained to make predictions or classify new data based on a labeled dataset. In unsupervised learning, the algorithm receives input data and gradually learns to identify correlations, similarities, and differences among the data until it derives recurring patterns and structures. In the case of reinforced learning, the process is typically focused on a series of decisions and based on a reward system when the desired decision is taken.

AI-based systems underpin such diverse software as speech recognition and computer vision, machine translation, search engines, virtual assistants, and so on<sup>1</sup>. Speech recognition models have traditionally been based on hidden Markov models (HMMs) and, before that, on dynamic time warping (DTW), but in recent years such models have been supplemented with recurrent neural networks (RNNs). Computer vision is a macro field that encompasses (among others) scene reconstruction, object detection, facial recognition, image restoration, and so on. The systems on which it is based are different according to the different purposes: image classification generally rests on convolutional neural networks (CNNs), deep learning algorithms inspired by the organic visual cortex, and designed to automatically and adaptively infer patterns and features from an image. Different types of computer vision may aim at detecting objects (using region-based convolutional networks, or R-CNN) or techniques such as You Only Look Once (YOLO), which allow objects to be detected and tracked in real time. Machine translation utilizes natural language processing (NLP) techniques<sup>2</sup>, which have long been based on statistical models and algorithms (statistical machine translation, or SMT), rule-based grammatical systems (rule-based machine translation, or RBMT), or a combination of the two (hybrid machine translation, or HMT), but are themselves increasingly centered on neural networks (neural machine translation, or NMT). NMT is nowadays at the core of the most

---

<sup>1</sup> For pragmatic reasons I keep out of this list many other automations, often combining several of the techniques, processes and models described here – for example, I will not discuss advanced driver-assistance systems (ADAS), nor social media algorithms and the like.

<sup>2</sup> Natural Language Processing is a subfield of artificial intelligence and computational linguistics focusing on the interaction between computers and human language. NLP can be divided into Natural Language Understanding (NLU), linked to machines’ ability to understand the meaning of human language, and Natural Language Generation (NLG), involving the artificial creation of meaningful sentences.

popular translation software and typically relies on recurrent neural networks (RNNs) - more specifically, a variant known as the sequence-to-sequence (Seq2Seq) model – or more complex transformer architectures, deep learning models specifically designed for sequence-to-sequence tasks (it is the case of DeepL). While both are based on an encoder/decoder model, the latter are more complex as they incorporate self-attention mechanisms, which enable them to process texts in a non-sequential manner by weighing the importance of different words within sentences. Traditional search engines similarly leverage natural language processing techniques to understand user queries and improve search accuracy, and either machine learning or deep learning to optimize search results based on user preferences and behaviors (an example is Google Search, and a clear overview of the systems that currently underlie the search mechanism is given by Google itself [1]). Virtual assistants can, again, be very different systems. Some are predominantly visual (e.g., Google Lens, Samsung’s Bixby Vision), and combine computer vision techniques based on convolutional neural networks, optical character recognition (OCR), and machine translation to recognize and interact with observed objects and situations. Many virtual assistants incorporate chatbots, themselves based on natural language processing and (usually) machine learning; in case they are based on voice commands (e.g., Siri, Alexa), such software also incorporates speech recognition. More recent examples of chatbots are OpenAI’s ChatGPT, Microsoft’s Bing Chat (which relies on OpenAI’s GPT-4), and Google’s Bard, all examples of the latest generative artificial intelligence. Apart from several differences in their training, the GPT (Generative Pre-trained Transformer) model and the LaMDA (Language Model for Dialogue Applications) one, on which Bard is based, are both based on the transformer architecture. The key feature of GPT is its autoregressive nature, which means that it generates text one step at a time, based on previously generated content. It does so by predicting the probability distribution over the entire vocabulary for each word position and following on the previously generated words, incorporating randomness to create diversity in the outputs. Bard, instead – which, at the time of writing this paper, is still not publicly available in my country - is a bidirectional transformer model, which means it considers both right and left context when making predictions.

Generative AIs, due to their unprecedented affordances, are indeed what, in the last months, has subsumed, in the mainstream debate, the whole concept of AI, with particular emphasis on text generation (the aforementioned chatbots), music composition (which will not be discussed here), and image generation. This came with a few overstatements, such as the claim of their occurred sentience [2] or the recent call to radically halt their development [3], in a *Time* op-ed that veers into apocalypticism the more reasonable and urgent appeals to regulate a complex subject matter [4] that currently eludes several of the existing legislative apparatuses - first and foremost, in terms of copyright, but also because of the massive social and economic repercussions that widespread use of generative AI can entail [6]. Nonetheless, these are still considered ‘Weak AI’ or ‘Narrow AI’, ‘Strong AI’ beginning with Artificial General Intelligence (AGI). In this sense, the claims about the alleged sentient nature of LaMDA should rather give pause to the need for more up-to-date tools than the Turing test. While the consequences in terms of labor are probably the most discussed in the debate about text-based generative AI (which is already integrated into the workflow of various professional environments, from cultural work to programming) [9], attention is also drawn to a potential loss of referentiality [7] - an extremely sensitive issue in an era that already faces post-truths and deepfakes [8] - those regarding AIs used for image generation have so far mainly regarded the topics of creativity and related copyright issues.

While this paper focuses on creativity, doing so requires a preliminary contextualization of the mechanisms by which currently available AIs generate images. Presently, the architectures used for image generation are of three main types. The first is generative adversarial networks (GANs), which consist of two neural networks (a generator and a discriminator) trained to compete. The generator generates new data samples similar to existing images from a given dataset, while the discriminator tries to distinguish between the generated and real patterns. The two networks are trained simultaneously: the generator attempts to deceive the discriminator and the discriminator tries to correctly identify the real samples. Those that pass the discriminator’s sieve are the final outputs. The second type is the diffusion probabilistic model (DPM), or diffusion model (diffusion models can perform various tasks, including denoising, restoring, or enhancing the resolution of existing images, but I will focus here on image generation alone). Generating images involves progressively introducing noise to an initial image until gradually creating a new one. This controlled introduction of noise is facilitated by a diffusion neural network (hence the name). The third type – adopted by most recent generative AIs - relies on an evolution of this mechanism, known as ‘stable diffusion’, whose first phase is carried on by a transformer model, and consists of image encoding through a neural network pre-trained on a large-scale dataset containing (billions of) image-text pairs, hence capable of associating images and their corresponding textual descriptions. The embeddings it creates are combined to form a joint representation of image and text, capturing the semantic meaning of both. This can be used for various tasks, such as image and text classification, object detection, and so on [10]. In stable diffusion models, further passages convert a text prompt into a text and then an image embedding, which the diffusion neural network transforms into images.



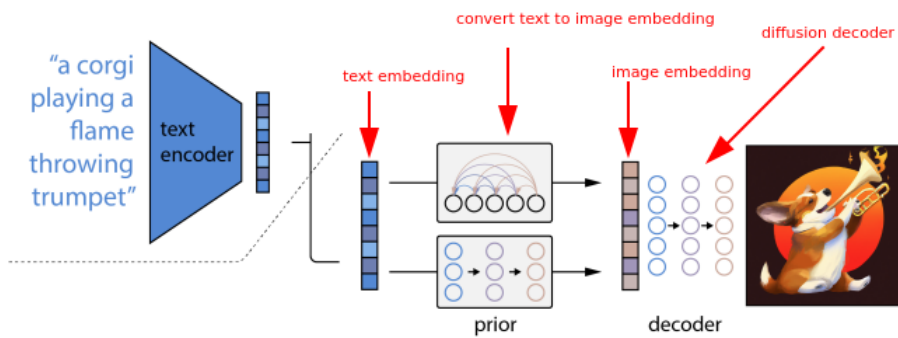


Figure 1. DALL-E2's image generative process [11].

This architecture saw a whirlwind acceleration in 2022, being at the base of the most known current Text-to-Image (TTI) generative models: DALL-E2, Midjourney, and Stable Diffusion itself. DALL-E (whose name is a portmanteau of WALL-E, the robot from the Pixar cartoon of the same name, and Surrealist painter Salvador Dali) debuted in January 2021, originally relying on GANs, but migrated to a stable diffusion model in September 2022 with DALL-E2 (currently asking to pay a small fee per image). Stable Diffusion is an open-source model released by the start-up company Stability AI in August 2022, and whose latest release is version 2.1 (free). Midjourney was released through Discord in July 2022 by the independent research lab Midjourney, Inc, and is currently in version 5.1 (it requires a monthly subscription). Apart from the prices, the main differences between the three concern the specifics of their processes and training. In fact, there are some differences in the way the process of encoding and decoding takes place in these models: Stable Diffusion relies on a succession of denoising autoencoders (through a variational autoencoder, or VAE, and the U-Net convolutional neural network developed by the University of Freiburg); DALL-E2's process begins with the CLIP (Contrastive Language-Image Pre-Training) neural network, created by OpenAI, and ends by upscaling the image generated by the diffusion model through another convolutional neural network. Midjourney has been very elusive about both their training and their model, and it has been speculated that it is based on a GAN (StyleGAN2: [12]) or a diffusion model quite similar to the one implemented by Stable Diffusion [13]. Concerning the datasets on which they were trained, DALL-E2 and Midjourney have not disclosed theirs, but we know that Stable Diffusion mainly relies on the LAION dataset, created by a non-profit organization called LAION, that processes through a CLIP model image-text pairs obtained from another non-profit organization called Common Crawl. Common Crawl, in turn, gets its images by scraping billions of web pages monthly and releasing them as openly available datasets [14]. Although, then, this is a transformative process, where training images data are not stored (vectors are), each isolated source is weighted (to the extent that significant changes in the datasets are needed to alter an output consistently and significantly), and singular outputs are subject to a substantial degree of randomness, generative AIs are uncannily better than the average human being at replicating the style of specific authors or artworks. The rest of the article will discuss this process of appropriation and re-use of existing material.

## 2. GENERATIVE ART, CREATIVITY, AND SHARED MEMORY

While generative art predates AIs [14, 15], it has certainly experienced a dizzying acceleration in recent years. In 2015, Google launched DeepDream, a convolutional neural network that found and enhanced patterns in images (a process known as algorithmic pareidolia), thus resulting in deliberately over-processed outputs. Soon, numerous apps adopted the mechanisms behind DeepDream to transform one's photos into images in the style of existing artists and works. In 2018, Paris-based art collective Obvious created and printed on canvas Edmond de Belamy, a generative adversarial network portrait, which achieved great notoriety after being auctioned at Christie's for \$432,500. In 2021, the first version of DALL-E was disclosed by OpenAI. In August 2022, Jason Allen won the Colorado State Fair's art competition in the category "emerging digital artists", with an artwork titled "Théâtre d'opéra spatial", created using Midjourney, which gave rise to heated controversy. Around the same period, the aforementioned generative AIs opened up to the general public.

The use of generative AIs surely has the potential to automate certain aspects of artistic creation, reducing the need for (some kinds of) human labor and transforming the field. Naturally, this raises several concerns, including ethical issues related to non-factual references (or 'hallucinations': [16]), biases incorporated from the datasets [17], repercussions on the labor market [18], intellectual property rights, and fair compensation for human creators [19]. Debates in the visual arts community have arisen about how AIs rely on datasets acquired by circumventing - according to some, contravening - the traditional ways of copyright [20, 21]. Without a doubt, generative AIs act as a function of an immense shared memory, made of the large image datasets used to train their neural networks. Nonetheless, it is controversial whether this process

exploits the authors of the images that are part of their datasets. Some argue that using copyrighted material without permission is a violation of intellectual property rights, while others point out that the training datasets consist of images that are freely available on the internet, and the generated images are not exact replicas, making it difficult to determine who the original influences are and whether they are being exploited: “AI art models do not store images at all, but rather mathematical representations of patterns collected from these images. The software does not piece together bits of images in the form of a collage, either, but creates pictures from scratch based on these mathematical representations” [20]. Overall, it is clear that the use of generative AIs raises complex ethical questions about authorship, intellectual property, and the role of technology in the creative process.

Nonetheless, the use of AIs in art creation, which some oppose to individual creativity as the manifestation of the capacity of a single genius, rather than rupturing with existing creative practices, points to a new phase of relying on what I define the “shared memory” that has always marked art creation, a process not far from what Perloff has defined “unoriginal genius” [22] and Goldsmith “uncreative writing” [23]. In this sense, the re-use of existing works is not simply a matter of imitation or replication, but instead a complex interplay between the artist’s own creativity and the styles, ideas, and techniques they have absorbed from others. Indeed, artistic originality as a defining characteristic of creativity is a relatively recent construct. Before the modern era, artists were not expected to create something entirely new, but to imitate and build on existing models. The idea of the individual artist as a genius producing unique works of art is a product of the Romantic era, which placed emphasis on personal expression and emotion (from an artistic point of view) and needed clear guidelines to attribute the paternity of a work of art (from a more pragmatic, industrial perspective).

However, artistic creativity has always relied on existing works, authors, and schools to create new art. According to Margaret Boden’s influential taxonomy [15], there are three types of creativity: combinatorial, which involves generating unfamiliar combinations of familiar ideas; exploratory, which relies on existing stylistic rules or conventions to generate novel structures; and transformational, which alters some defining dimension of what Boden calls ‘the conceptual space’ that previously existed. Although Boden’s categories are easily misinterpreted, and even though they are focused on identifying different types and strategies of creative processes, it is significant that they all presuppose that creativity lies in what is already existing. Artists draw inspiration from their predecessors, learn from their techniques, and build on their ideas. The unfolding of art history can as such be seen as a continuous dialogue between the past and the present, with each new movement or style emerging as a response to what came before it, and every single work as carrying, both consciously and unconsciously, the inscription of the whole system of references held by the single artist. The artist’s relationship with shared artistic memory is not always direct or deliberate: sometimes, artists are influenced by works of art that they have never seen, but have known, heard, or learned about through cultural osmosis – for example, they may have seen a pastiche, a parody, or a reference. Other times, artists may be unconsciously influenced by broader social and cultural trends. Again, influences may depend on the sociocultural milieu in which artists are immersed – with mass communication, the continuous flow of information that we process every day and are inevitably distilled in artistic production.

Although the idea of art with which we were raised was that of the product of an individual, original, inspired mind, everything in the history of art points to the exact opposite. Art creation has always been at the crossroad between individual and collective style, originality, and reuse. Twentieth-century art largely acknowledged it through techniques such as ready-made, collage, assemblage, cut-up, sampling, conceptual art. Away from avant-gardes, this has been particularly evident in a ‘dirtier’, openly industrialized art practice such as comics, where copying other people’s works was initially born out of necessity as a time-saving, barely licit practice in the creation of cheap, pulp stories. Not accidentally, now that comics have acquired a more consistent cultural capital, it is, on the contrary, used mainly by recognized authors as a way to pay homage to artists they admire and to ground their production in the system of existing works through intertextual references [24].

In this sense, creation employing AI makes no exception. Rather than a novum, the process can be seen as a continuation and a re-proposition of mechanisms that have always underpinned creativity [25, 26], albeit with a new set of tools and techniques. As said, every output obtained by generative AIs carries within it the potential inscription of *all* the previously existing works through which it was trained, drawing attention to how creativity has always been about reusing and recombining existing ideas and material; yet, at the same time, it results from *none* of them specifically, because the encoding/decoding process results in a deviation that it is tempting to attribute to a sort of synthetic unconscious which would take a step further from Hayles’ idea of ‘nonconscious cognition’ [27, 28] (after all, part of the behavior shown by LLMs is still inexplicable to their creators: [29, 30]). This introduces a new non-human element that situates between an actor and a tool, performing part of the creative labor in a way that, at present, we can only partially deconstruct and control. The novelty of the technological unconscious lies not in the process per se, but in the fact that AIs partake in the memorial act that substantiates artistic creation in a way that a chisel or a brush couldn’t. This thus calls both for a reframing of creativity as a collective practice that relies first and foremost on memory, and of the agency behind creation as a way more

ramified process than proclaimed by the modern, Western, Romantic idea of genius, one that encompasses both human and non-human actors in a shared creative venture.

Nonetheless, to provide effective results, generative AIs currently need a careful, ad-hoc prompting (in the case of TTIs) that is all but intuitive, instead being somewhere in between a new narrative form and a technical language; current AIs still struggle with offering complete control over their images, especially when asked for logical permutations, specific quantities, shapes or spatial requirements, qualities, and consistency over a set of images. Similarly, GANs' training entails selecting the datasets, which implies choices and directions imparted by the human authors controlling the process. Furthermore, the output image obtained from a generative AI not only is the result of significant tweaking, but usually needs to be selected, retouched, and edited, before reaching its ultimate form: if generative AIs are not as simple a tool as a chisel, neither they are single authors: there is no doubt that the results one can obtain through them is, up to now, always a co-creation between a human and a non-human actor, where the former still plays the most significant role.

### 3. COMICS, THE DIGITAL, AND THE EMERGENCE OF AIS

Comics are a particularly interesting test case in this regard, for a few different reasons. First of all, because of the close link they have been able to weave with digital culture, from which they have been interpenetrated and to which they have perhaps lent some systems of signification (the language of memes, for example, owes a great deal to that of comics).

The last two decades have seen a significant shift towards the digitization of comics creation. Comics authors have progressively abandoned traditional tools like pencils, inks, and paper, in favor of digital tools like tablets and software that allow faster, more easily editable works. Working only with analog tools is now a non-prototypical choice for comics authors, as would be writing a book with a typewriter; this kind of creative choices is generally meant to make up for (the perception of) the lack of authenticity and aura of the digital. Digitization has also changed the way comics are distributed and consumed. Most print-based comics now know a double life in their digital form, born-digital comics have explored the affordances of their supports and experimented with new forms of storytelling, while some comics platforms (e.g., Comixology, Webtoon) have acquired a widespread diffusion and comics circulating on social media platforms have subtly remediated the characteristics of the comics form in light of the increased, accelerated communication with the readers thanks to the digital environment. The whole process of progressive digitalization has come with some discontent, and the initial assertions by many scholars and practitioners that digital comics – remarkably, those that transgress the static nature of the image and the division into pages and panels - are not *proper* comics [31]. This is one of the reasons why many digital comics aim at an afterlife in print form where their authors can capitalize on their online success through the traditional market, still the locus of both proper cultural legitimation and commercial success.

Secondly, the relevance of comics lies in their hybrid, 'dirty' nature, in perpetual oscillation between culturally legitimized artistic practices and fully commodified industrial production apparatuses. As a consequence, their symbolic structures are always conflicted between an industrial side relying on supply chains and an artistic one that reveres the myth of the author. Unsurprisingly, then, several comics authors have been on the frontline in voicing their concerns against AI-created comics [32, 33], clamoring for greater protections in terms of transparency of the datasets on which such AIs are trained, the right to opt out of such corpora, and payment of royalties if their works are part of the material on which this training is conducted. However, such claims, even if they seem to have been provisionally accepted by the European legislator [34] (less so, though, by the US court [35]), open the way to much broader questions: how to regulate copyright compliance not only in the US or Europe, but from a supranational perspective? What kind of compensation can result from being a billionth of the training corpus from which a generated work is derived? Would the resulting system be sustainable, or would it end up disrupting the profits of small- and medium-sized artists, as it has already happened in music through the mechanism of streaming platforms (despite them having a more direct royalties mechanism)?

Be as it may, I argue that the changes introduced by AIs are not limited to the most evident repercussions affecting the Taylorist side of comics production (the comics studios where writers, drawers, colorists, inkers, and editors are separated persons that act as components of an assembly line), but extend to its more *artistic* embodiment, regarding independent, *auteur* comics and graphic novels<sup>3</sup>. It is further interesting, in this sense, that generative AIs are currently more efficient in making paintings or photorealistic images than in producing comics, which makes it still a liminal practice - although this may also depend on the extent and variety of datasets containing comics and the generative AI models used so far.

However, and although we are still far from a hegemony of comics made this way, 2022 has indeed seen the emergence of a first batch of AI-created comics and graphic novels, the most notable being *Fastwalkers* by Ilan Manouach; *The Abolition of Man*, by Carson Grubaugh; *Prompt. Conversations with AI*, by Dave McKean; and *Lovebot*, hosted in the Italian self-

---

<sup>3</sup> I will stick to this dichotomy for the sake of the brevity of this article, but it should be nuanced and deconstructed [36, 37].

produced and self-distributed comics magazine Čapek, and attributed to an AI aptly called Nobot (but whose human authors are probably Jazz Manciola and Hurricane Ivan, both members of the Čapek editorial staff).

Despite having been all published in 2022, the four stem from different paths and practices in drawing with AIs: Manouach, who is a conceptual comics artist, had already been experimenting with AI creation in shorter forms, with the *Neural Yorker* – a series of automatically generated, New Yorker-ish strips (2020) – and with the conceptual/erotic graphic novel *Le VTT comme je l'aime* (2022), before devoting himself to *Fastwalkers*. The latter is a 512-pages manga generated through a GAN; the text of the manga was, in turn, generated by an artificial intelligence model called GPT-3, and a random generator arranged the layout. McKean is (mainly) a comics artist known for his unique, innovative approach to visual storytelling, blending various artistic mediums such as illustration, painting, photography, sculpture, and digital techniques. His *Prompt. Conversations with AI* was realized in a very short period (less than two weeks) by experimenting on Midjourney, and explores the role of artificial intelligence in the creative process. Grubaugh is a painter and an art professor; his evocative title plays with the machinic imaginary, but actually comes from a philosophical essay written by C.S. Lewis in 1943, dealing with education and morality, that Grubaugh adapted into a graphic novel with Midjourney. Finally, all we know about *Lovebot* is that it is a short collective comic that appeared on the Italian underground webzine Čapek, and was allegedly created by training a GAN on a corpus of erotic comics.

Two of these works have thus been created via TTIs, two by processing a curated database of stylistically coherent comics. Two of them present, more or less explicitly, Midjourney as a co-creator, while the other two frame the AI as a single author, to the point, in the case of *Lovebot*, of hiding the name of the human (co-)authors (Manouach's name is, on the contrary, clearly visible on the cover, but this has to do with the commercial side of the authorial function. Moreover, two of the four works are graphic novels; one is a manga – although by an avant-garde comics artist, not a mangaka; and one is an experimental underground comic. In all cases, the creators – a conceptual artist, an experimental storyteller, a painter, and an underground collective – are backed up by a consistent cultural (sometimes counter-cultural) capital, which positions their works no matter their specific features. If there is a formal or structural trait that holds together the four works mentioned here, it is their eccentric position in relation to prototypical comics, followed by the symbolic insistence on the uncanny charge related as much to the erotic component (an alienated and dehumanized eroticism) as to that of the imaginary of the human-like machine. While the two comics made with Midjourney stand out for the markedly pictorial nature of their trait, the works created with the GANs return imperfect, unnatural, segmented bodies fused in impossible poses. *Fastwalkers*, which is the work that most pushes this idea to its radical consequences by also entrusting an AI with the verbal component of its narrative (and, thus, with its plot) is, in fact, an unreadable text, whose value lies in its reflection on the production practices and semiotic conventions of comics, but which does not establish solid causal links in its narrative progression, preventing from any kind of linear reading or search for an exact meaning. In this, one wonders how long such works will be experimental and subversive, and how far they align themselves, instead, with certain ongoing trends in digital culture (e.g., glitch art, post-ironic memes) in which referentiality and meaning are constantly hampered by a frayed and fragmentary signification, the reading of which offers no closure apart for the realization of an inescapable, fundamental lack of meaning. Digital culture often foregrounds humorous works, and humor – like creativity – arises from a deviation, a cognitive dissonance stemming from what already exists. Maybe, then, works of art created through AIs are not only a reproduction, in a more evident form, of the series of reuse and remix processes that have been a constitutive part of artistic creation since time immemorial and that after a downturn during modernity have regained, during the 20th and 21st centuries, an increasing preponderance. Maybe, even more significantly, they express and allow for a sensation of loosening of meaning common to much contemporary sensibility.

#### 4. CONCLUSIONS

The emergence of AIs in comics creation has opened up new opportunities and possibilities, but also new criticalities. While AI-generated art is often seen as a departure from existing artistic practices, it is instead profoundly connected to the history of art, foregrounding creative processes grounded on the kind of shared memory that underlies traditional artistic creation. As such, the process of synthetic artistic generation is not a replacement for human creativity but rather a complement to and a continuation of it, albeit in a new form. The use of AI for artistic creation implies an unprecedented distribution of cognitive processes, but one that still needs guidance and whose randomness is still significant. The example of comics shows how the impact of image-generating AIs involves practical repercussions on the industrial aspects as much as theoretical implications for contemporary creation. Ultimately, the current use of AI in comics creation and artistic production at large, and its potential consequences and ramifications, are still being explored. It is up to all of us to grapple with the issues they will face us with, and determine how to best navigate this frontier.

## REFERENCES

- [1] Google. “How AI Powers Great Search Results,” February 3, 2022. <https://blog.google/products/search/how-ai-powers-great-search-results/>
- [2] Lemoine, Blake. “Is LaMDA Sentient? — an Interview.” Medium, June 11, 2022. <https://cajundiscordian.medium.com/is-lambda-sentient-an-interview-ca64d916d917>
- [3] Yudkowsky, Eliezer. “The Open Letter on AI Doesn’t Go Far Enough.” Time, March 29, 2023. <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>
- [4] Future of Life Institute. “Pause Giant AI Experiments: An Open Letter.” Future of Life Institute, March 22, 2023. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
- [5] Glover, Ellen. “AI and Copyright Law: What We Know | Built In.” builtin.com, April 18, 2023. <https://builtin.com/artificial-intelligence/ai-copyright>
- [6] Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. “GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.” ArXiv:2303.10130 [Cs, Econ, Q-Fin], March 17, 2023. <https://arxiv.org/abs/2303.10130>
- [7] Hsu, Tiffany, and Steven Lee Myers. “Can We No Longer Believe Anything We See?” The New York Times, April 8, 2023, sec. Business. <https://www.nytimes.com/2023/04/08/business/media/ai-generated-images.html>
- [8] Wong, Matteo. “We Haven’t Seen the Worst of Fake News.” The Atlantic, December 20, 2022. <https://www.theatlantic.com/technology/archive/2022/12/deepfake-synthetic-media-technology-rise-disinformation/672519/>
- [9] Zarifhonarvar, Ali. “Economics of ChatGPT: A Labor Market View on the Occupational Impact of Artificial Intelligence.” papers.ssrn.com. Rochester, NY, February 7, 2023. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4350925](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4350925)
- [10] Jung, Col. “Generative AI Art — DALL-E, Midjourney & Stable Diffusion Explained.” Medium, May 13, 2023. <https://generativeai.pub/generative-ai-art-the-road-to-dall-e-midjourney-stable-diffusion-3b3219d97f02>
- [11] Kosar, Vaclav. “OpenAI’s DALL-E 2 and DALL-E 1 Explained.” vaclavkosar.com, April 13, 2022. <https://vaclavkosar.com/ml/openai-dall-e-2-and-dall-e-1>
- [12] ILLA Blog. “DALL-E vs. Midjourney AI vs. Stable Diffusion,” May 12, 2023. <https://blog.illacloud.com/dall-e-vs-midjourney-ai-vs-stable-diffusion-a-comparison-of-ai-models-that-can-generate-images-from-text/>
- [13] Nerds Chalk. “Does Midjourney Use Stable Diffusion?,” April 7, 2023. <https://blog.illacloud.com/dall-e-vs-midjourney-ai-vs-stable-diffusion-a-comparison-of-ai-models-that-can-generate-images-from-text/>
- [14] Ashe, Matthew. “Will AI Text-To-Image Tools Replace Artists or Empower Them?” euronews, December 30, 2022. <https://www.euronews.com/next/2022/12/30/dalle-2-stable-diffusion-midjourney-how-do-ai-art-generators-work-and-should-artists-fear->
- [15] Boden, Margaret A. *Creativity and Art*. Oxford University Press, 2012.
- [16] Weise, Karen, and Cade Metz. “When A.I. Chatbots Hallucinate.” The New York Times, May 1, 2023, sec. Business. <https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html>
- [17] Offert, Fabian, and Thao Phan. “A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space.” ArXiv:2211.06323 [Cs], October 26, 2022. <https://arxiv.org/abs/2211.06323>
- [18] Baschuk, Bryce. “A.I. Will Cause ‘Significant Labor-Market Disruption’ over next 5 Years, Says World Economic Forum.” Fortune, May 1, 2023. <https://fortune.com/2023/04/30/ai-disruption-jobs-labor-world-economic-forum-report/>
- [19] Vallance, Chris. “‘Art Is Dead Dude’ - the Rise of the AI Artists Stirs Debate.” BBC News, September 13, 2022, sec. Technology. <https://www.bbc.com/news/technology-62788725>
- [20] Vincent, James. “AI Art Tools Stable Diffusion and Midjourney Targeted with Copyright Lawsuit.” The Verge, January 16, 2023. <https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart>
- [21] ———. “Getty Images Is Suing the Creators of AI Art Tool Stable Diffusion for Scraping Its Content.” The Verge, January 17, 2023. <https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit>
- [22] Perloff, Marjorie. *Unoriginal Genius*. University of Chicago Press, 2010.
- [23] Goldsmith, Kenneth. *Uncreative Writing : Managing Language in the Digital Age*. New York: Columbia University Press, 2011.
- [24] Crucifix, Benoît. *Drawing from the Archives: Comics Memory in the Contemporary Graphic Novel*. Cambridge UP, 2023.
- [25] Kaufman, James C., and John Baer (Eds.). *Creativity across domains: Faces of the muse*. Psychology Press, 2005.
- [26] Kaufman, James C., and Sternberg, Robert J. (Eds.). *The Cambridge handbook of creativity*. Cambridge University Press, 2019.
- [27] Hayles, N. Katherine. *Unthought. The Power of the Cognitive Nonconscious*. University of Chicago Press, 2017.
- [28] Hayles, N. Katherine. *Postprint: Books and becoming computational*. Columbia University Press, 2021.
- [29] Bowman, Samuel R. “Eight things to know about large language models.” ArXiv:2304.00612, April 2, 2023. <https://arxiv.org/abs/2304.00612>
- [30] Cassella, Carly. “A Nightmare Face Is Haunting AI Art, and There’s a Reason We Shouldn’t Look Away.” Science Alert, December 24, 2022. <https://www.sciencealert.com/a-nightmare-face-is-haunting-ai-art-and-theres-a-reason-we-shouldnt-look-away>

- [31] Busi Rizzi, Giorgio. «Digital comics: an old/new form». Maaheen Ahmed, ed. The Cambridge Companion to Comics. Cambridge UP, 2023.
- [32] Arcidiacono, Francesco. “Help Protect Our Art and Data from AI Companies.” gofundme.com, May 12, 2023. <https://www.gofundme.com/f/help-protect-our-art-and-data-from-ai-companies>
- [33] EG AIR (European Guild for Artificial Intelligence Regulation), “Help Protect Our Art and Data from AI Companies.” Egair, , April 11, 2023. <https://www.egair.eu/>
- [34] Vincent, James. “Reported EU Legislation to Disclose AI Training Data Could Trigger Copyright Lawsuits.” The Verge, April 28, 2023. <https://www.theverge.com/2023/4/28/23702437/eu-ai-act-disclose-copyright-training-data-report>
- [35] Novak, Matt. “AI-Created Images Aren’t Protected by Copyright Law according to U.S. Copyright Office.” Forbes, February 22, 2023. <https://www.forbes.com/sites/mattnovak/2023/02/22/ai-created-images-in-new-comic-book-arent-protected-by-copyright-law-according-to-us-copyright-office/>
- [36] Baetens, Jan, and Hugo Frey. The Graphic Novel: An Introduction. Cambridge UP, 2015.
- [37] Beaty, Bart. Comics Versus Art. University of Toronto Press, 2012.

# ChatGPT: un Pappagallo Stocastico può essere di aiuto a un Vero Ricercatore (Umanistico)?

Fabio Ciotti<sup>1</sup>

<sup>1</sup> Università degli Studi di Roma Tor Vergata, Italia – fabio.ciotti@uniroma2.it

## ABSTRACT

Da quando è stato rilasciato da OpenAI nello scorso autunno il sistema di interazione linguistica ChatGPT, basato a sua volta sul modello linguistico GPT-3.5, è assurdo agli onori della cronaca e del dibattito, sia specialistico sia pubblicistico. In questo paper, dopo averne esaminato i principi, intendo esaminare le possibili ricadute di questa classe di sistemi sulle pratiche della ricerca scientifica in ambito umanistico e in particolare in ambito umanistico digitale o computazionale. La tesi che propongo è che ChatGPT già nello stato di avanzamento in cui si trova oggi 1) rappresenta un ottimo strumento di supporto per la progettazione e implementazione di *workflow di ricerca*; 2) potrebbe avere un impatto rilevante nella qualità e nella efficacia della ricerca umanistica digitale/computazionale.

## PAROLE CHIAVE

Intelligenza artificiale (AI); Large Language Models (LLM); ChatGPT; research workflow design; metodologia della ricerca; metodologia delle DH; explainable AI

## 1. INTRODUZIONE

Da quando è stato rilasciato da OpenAI nell'autunno 2022 il sistema di interazione linguistica ChatGPT [1], basato a sua volta sulla famiglia di modelli linguistici (LLM) generativi GPT<sup>1</sup>, è assurdo agli onori della cronaca e del dibattito, sia di quello specialistico sia di quello della pubblicistica non specializzata e del chiacchiericcio dei social media. Tra i molti temi sui cui si è incentrato tale dibattito, che non è possibile riassumere in questa sede, non mancano quelli relativi alle applicazioni degli LLM generativi e delle relative conseguenze, nel mondo della ricerca e della didattica. Queste ultime in particolare hanno destato una notevole apprensione, almeno in parte giustificata poiché è evidente che un sistema di generazione linguistica così efficiente ed efficace sia nella coerenza sintattica sia in quella semantica dei testi generati, e dotato di una interfaccia semplice da usare, potrebbe sicuramente essere utilizzato dagli studenti per produrre almeno parte di quella serie di contributi scritti mediante i quali vengono giudicati e valutati [3]. Ma anche il settore dell'editoria scientifica è stato agitato da allarmi che hanno portato alcuni importanti editori e repository di *preprints* a definire dei veri e propri documenti di policy [4] [5], nonché la MLA a fornire una prima indicazione sulle modalità di citazione del contributo di un LLM al contenuto di un lavoro scientifico [6].

Mi interessa piuttosto in questa sede riflettere sulle possibili ricadute di un sistema come questo – o dei suoi consimili che a breve inonderanno il mercato della cosiddetta intelligenza artificiale –, sulle pratiche della ricerca scientifica in ambito umanistico e in particolare in ambito umanistico digitale o computazionale, ciò di cui ci occupiamo tutti noi che ci riconosciamo come membri della comunità dell'associazione di informatica umanistica italiana. Intendo dunque fornire alcuni spunti di riflessione relativamente al fatto che ChatGPT già nello stato di avanzamento in cui si trova oggi:

1. rappresenta un ottimo strumento di supporto per la progettazione e implementazione di *workflow di ricerca*;
2. potrebbe avere un impatto rilevante nella qualità e nella efficacia della ricerca umanistica digitale/computazionale.

Queste due osservazioni valgono in primo luogo se è verificata la condizione necessaria che studenti e studiosi di area umanistica abbiano una sufficiente conoscenza del modo in cui un LLM come GPT funziona, come è vero per ogni altro metodo matematico/computazionale [7, Par. The Missing Curricular Foundation], [8].

---

<sup>1</sup> Nel momento in cui scriviamo il servizio web ChatGPT può usare due diverse versioni della generazione 3.5 e la più recente ed estesa, ancorché assai meno trasparente, generazione 4 del modello [2].



## 2. COSA È E COME FUNZIONA CHATGPT, IN BREVE

ChatGPT è, come detto, un modello linguistico conversazionale basato sui modelli linguistici GPT, acronimo di *Generative Pre-trained Transformer*. Questi a loro volta sono modelli probabilistici del linguaggio naturale costituiti da enormi reti neurali artificiali (ANN) basate su una complessa architettura detta a *transformer con attenzione* [9]. Obiettivo di questo genere di sistemi, di cui GPT è solo il più noto al grande pubblico, è generare una risposta linguistica (anche in diverse lingue) a un input di testo (*prompt*) in modo linguisticamente coerente (sia dal punto di vista sintattico sia da quello semantico) e auspicabilmente pertinente.

I modelli GPT sono reti neurali composte da un'enorme quantità di parametri (i pesi assegnati alle connessioni tra i nodi, gli atomi della conoscenza acquisita da una ANN): il modello più vasto della generazione 3 conta 175 miliardi di parametri [10]; purtroppo per la generazione 4 OpenAI ha assunto una policy di assoluta chiusura per cui non sono noti dettagli sia architetturali sia sulle fonti utilizzate per l'addestramento del modello. Il suo funzionamento è basato su una combinazione di stack paralleli composti a loro volta da diversi strati di moduli *transformer* (dei codificatori di vettori di numeri reali) detti *decoders* con *multihead self-attention*. In fase di generazione del testo ogni token del prompt di input viene elaborato in modo indipendente e il modello utilizza l'attenzione per capire quali token del contesto sono rilevanti per la comprensione del significato del token. Per la produzione dell'output finale GPT utilizza un processo di decodifica per selezionare la parola successiva da generare in base alla probabilità che segua la sequenza di parole esistenti nel prompt. Questo processo viene ripetuto più volte, fino a quando viene generato un testo completo. Si noti che la scelta finale di ogni parola generata è soggetta a un algoritmo stocastico per evitare che sia sempre la parola *più* probabile a essere scelta. Il modello, si potrebbe dire, combina la sua conoscenza del linguaggio con la sua capacità di valutare la probabilità di ogni parola in base al contesto in cui si trova unita a una certa dose di casualità [11].

Come GPT (e i vari altri modelli linguistici oggi disponibili) acquisisce la sua conoscenza del linguaggio? Il modello viene addestrato su una grande quantità di dataset testuali estratti dal Web, quali CommonCrawl ([commoncrawl.org](http://commoncrawl.org)), WebText2 (post su Reddit, [openwebtext2.readthedocs.io](http://openwebtext2.readthedocs.io)), Wikipedia e vari archivi di libri e pubblicazioni digitali. GPT-3 è stato addestrato con circa 400 miliardi di token, mentre come già detto nulla di preciso si sa della versione 4. L'addestramento, che richiede mesi di elaborazione con centinaia di unità GPU funzionanti in parallelo, si basa su un processo di auto-apprendimento automatico [12], a sua volta fondato sul noto algoritmo di *back-propagation*. Si tratta di un metodo matematico di ottimizzazione introdotto nello sviluppo di reti neurali sin dagli anni 80 dal gruppo *Parallel Distributed Processing* di Rumelhart, e McClelland [13]. L'input per l'addestramento di GPT è formalizzato sotto forma di sequenze di token, che vengono convertiti in numeri utilizzando un vocabolario che assegna poi a ogni token il suo *embedding vettoriale* [14]. Ad esempio<sup>2</sup>

1. Il - gatto - saltò - sul - tavolo
2. {2: "Il", 4: "gatto", 7: "saltò", 10: "sul", 12: "tavolo"}
3. 099837849898... - 934898689573589... - 093745282902...

In fase di addestramento, il modello riceve la sequenza di token [2, 4, 7, 10] come input e deve generare il token successivo [12]. In realtà inizialmente fornirà una risposta casuale – mettiamo “lago” – che viene valutata confrontando la distanza vettoriale tra la risposta del modello e la risposta corretta, fornita come parte del corpus di addestramento. La differenza tra la risposta del modello e la risposta corretta, definita “loss”, misura quanto il modello è lontano dalla risposta corretta e viene usata come input del processo di *back propagation*. Questo aggiorna i pesi (parametri) del modello utilizzando un algoritmo di ottimizzazione, che aggiusta gradualmente i pesi in base alla ‘perdita’ calcolata durante la fase precedente. Iterando questo processo di ottimizzazione durante l'addestramento, il modello continua a migliorare la sua precisione e a aumentare la sua capacità di generare un output coerente e pertinente che, nel caso dei modelli linguistici, significa individuare la (distribuzione di) probabilità che un insieme di parole ha di seguire una data sequenza testuale di input.

Ma la capacità conversazionale di ChatGPT, oltre alla fase di *self-supervised training*, è stata sviluppata adottando una combinazione di altre tecniche di *fine tuning* supervisionato [16]. In particolare, nell'addestramento di ChatGPT hanno avuto un ruolo importante una fase di addestramento guidato con esempi di etichettati dialogo (*instruction fine-tuning*) e una basata sul *reinforcement learning*. Il reinforcement learning è una tecnica di apprendimento automatico che consente a un modello di apprendere tramite la sperimentazione e il feedback. Nel contesto di ChatGPT, il modello riceve una ricompensa positiva se la risposta generata è rilevante e coerente con la conversazione in corso. Al contrario, riceve una ricompensa negativa se la risposta è errata o inappropriata. Questo feedback può essere utilizzato per modificare i pesi del

---

<sup>2</sup> L'esempio fornisce una visione semplificata della tokenizzazione di GPT, che in realtà non è basata su parole ma su unità sub-parola, individuate mediante l'algoritmo *Byte-Pair Encoding* (BPE), introdotto in [15]

modello, migliorando la sua capacità di generare risposte coerenti e appropriate in una conversazione. Ad esempio, se il modello genera una risposta che viene valutata come poco pertinente o inappropriata, i pesi della rete neurale possono essere modificati in modo da rendere meno probabile la generazione di risposte simili in futuro.

### 3. CHATGPT NEL DIGITAL LAB DELL'UMANISTA DIGITALE

Dopo avere riassunto i fondamenti teorici di questo genere di modelli neurali artificiali del linguaggio, torniamo al ruolo che essi possono assumere nelle pratiche di ricerca sperimentale che sono tipiche del campo umanistico digitale [17]. Ciò che intendo sostenere è che i sistemi linguistici generativi possono intervenire durante la progettazione del flusso del processo di ricerca successivo alla individuazione di una ipotesi di ricerca e di una serie di *research questions*. In particolare, mi riferisco alla fase di esplorazione nello spazio dei metodi e degli strumenti applicabili alle ipotesi di ricerca individuate dal ricercatore. Questo è dovuto al fatto che ChatGPT si è rivelato uno strumento particolarmente versato nella predisposizione di sistemi di prototipazione rapida che altrimenti richiederebbero molto tempo (sempre ammesso che il ricercatore abbia le necessarie competenze di sviluppo software) o il coinvolgimento nel gruppo di ricerca di studiosi di area tecnico-informatica già disponibili a partecipare al progetto stesso anche nella sua fase puramente esplorativa e aurorale. Ovviamente è ben lungi da me ritenere che, in mancanza di una adeguata conoscenza dettagliata del dominio di ricerca, delle possibilità offerte dai metodi computazionali e di una certa competenza almeno approssimativa nella redazione di programmi in un qualche linguaggio di programmazione o formalismo adeguato, sia possibile produrre alcunché di qualificato e significativo. Tuttavia, in presenza di queste condizioni i modelli conversazionali sono in grado di fornire

1. supporto alla creazione di workflow di ricerca
2. prototipazione software per il test dei metodi individuati di aiuto nella predisposizione
3. formalizzazione dei dati e dei dataset da utilizzare nella ricerca.

Ad esempio, un ricercatore di ambito letterario che volesse applicare tecniche di *machine learning* per analizzare grandi quantità di testi potrebbe utilizzare ChatGPT per ottenere indicazioni su come strutturare il workflow di analisi dei dati, quali tecniche di analisi testuale utilizzare, come preparare i dati per l'analisi, e come interpretare i risultati dell'analisi. Inoltre, ChatGPT può anche fornire indicazioni su come selezionare e predisporre i dati appropriati per l'analisi, (ad esempio quali fonti di dati utilizzare e come selezionare i dati in modo da limitare il rischio di *bias*). In definitiva un sistema come ChatGPT può essere considerato un *artificial research assistant* in grado di supportare studiosi le cui competenze di *data science* non sono professionali, nella fase di *operationalizzazione* dell'analisi. Quella che segue, ad esempio, è la trascrizione parziale di una interazione con ChatGPT volta a definire un workflow metodologico e a predisporre il codice Python necessario per l'individuazione di brani testuali che siano tematicamente attinenti al tema del lavoro<sup>3</sup>:

**F:** I would like you to help me design a workflow for the following task: I want to use Bert to detect passages of texts in a corpus of multilingual novels that are examples of a certain theme. For instance: I want to extract excerpts of text that are about a working activity. How could I proceed?

**GPT:** Sure, here's a suggested workflow for using BERT to detect passages of text in a corpus of multilingual novels:

Data preparation: Collect a corpus of multilingual novels and pre-process them to remove any noise and unwanted characters. You can also tokenize the text into sentences or paragraphs for easier processing.

Fine-tune a pre-trained BERT model: Fine-tune a pre-trained BERT model on a large corpus of text that includes examples of the theme you are interested in (in this case, working activity). You can use a pre-trained multilingual BERT model to handle the multilingual aspect of your corpus. The fine-tuning process involves training the BERT model on your specific task by providing it with examples of text that are about the theme you are interested in, and then fine-tuning the model to identify passages that match that theme.

Text classification: Use the fine-tuned BERT model to classify each sentence or paragraph in your corpus as either relevant to the theme or not. You can use a threshold for the model output to determine if a given sentence or paragraph is about the theme.

Extract relevant passages: Once you have identified the relevant sentences or paragraphs, you can extract the corresponding passages of text from the original corpus. You can combine adjacent sentences or paragraphs that are relevant to create longer passages.

---

<sup>3</sup> Questa interazione è stata condotta dall'autore nella preparazione di una ricerca presentata al seminario "Émotions, travail et récit" (Aix-en Provence, Marzo 2023) con il titolo "The 'feeling' of labor in Italian Literature 1800-2000: a distant reading approach".

**Post-processing:** Finally, you can apply any necessary post-processing to the extracted passages, such as removing duplicates or sorting by relevance.

Some additional considerations for this workflow include selecting an appropriate BERT architecture and hyperparameter tuning to optimize performance, as well as addressing any language-specific nuances in your corpus.

**F:** How many tagged examples are needed for fine tuning?

**GPT:** The number of tagged examples needed for fine-tuning a BERT model depends on several factors, such as the complexity of the task, the size and quality of the training data, and the performance requirements of the model. In general, the more examples you have, the better your model is likely to perform, as it has more information to learn from.

However, it's also important to keep in mind that fine-tuning a BERT model can be computationally expensive, so you may need to balance the number of training examples with the available computing resources and training time.

A general rule of thumb is to have at least a few hundred examples per class for good performance, but in some cases, thousands of examples may be needed for optimal results. In practice, you may need to experiment with different amounts of training data to determine the optimal amount for your specific task.

If you have a limited amount of labeled data, you can also consider using transfer learning to leverage a pre-trained BERT model's knowledge and fine-tune it with your smaller dataset. This can help improve the model's performance, even with limited labeled data.

**F:** Tell me if this can work use a fixed vocabulary of work-related terms to extract 200 random passages from my corpus; use them to fine tune a Bert model; then use the Bert model fine-tuned to detect passages in the rest of the corpus

**GPT:** Using a fixed vocabulary of work-related terms to extract random passages from your corpus can be a useful approach for creating a labeled dataset for fine-tuning a BERT model. However, there are a few potential issues with this approach:

**Limited diversity of passages:** By using a fixed vocabulary to extract passages, you may end up with a limited set of passages that all use similar or related terminology. This can limit the diversity of the training data and result in a model that is overly focused on specific keywords, rather than the broader theme of work-related activities.

**Bias in the labeled dataset:** Depending on how you select and extract the 200 passages, there may be some bias in the labeled dataset that could affect the performance of the model. For example, if the passages are all from a specific genre of novels or from a particular time period, the model may not perform as well on passages from different genres or time periods.

**Difficulty in scaling:** If you plan to use the fine-tuned BERT model to detect passages in the rest of the corpus, you may find that the fixed vocabulary approach does not scale well to larger or more diverse datasets. For example, there may be work-related terms or phrases that are not included in your fixed vocabulary, which could result in missed passages.

That being said, this approach could still work depending on the specifics of your corpus and the task you are trying to accomplish. If you are working with a smaller corpus and a well-defined set of work-related terms, using a fixed vocabulary to extract passages could be a good way to create a labeled dataset for fine-tuning a BERT model. However, if you are working with a larger or more diverse corpus, or if you want to detect a wider range of work-related passages, you may need to consider other approaches for creating your labeled dataset, such as using keyword extraction algorithms or manual annotation.

**F:** Can you write a python program to test my approach?

**GPT:** Sure! Here is some sample code in Python that demonstrates your approach for extracting work-related passages and fine-tuning a BERT model: [...]

#### 4. CONCLUSIONI E PROBLEMI APERTI

L'esempio che abbiamo fornito sopra (frutto di una attività di ricerca reale condotta dall'autore di questo contributo) mostra come i sistemi LLM, con tutti i loro limiti, se opportunamente e consapevolmente utilizzati sono di grande utilità nella fase di definizione e prototipazione rapida della ricerca, e permettono di automatizzare e documentare parti rilevanti del workflow (con ricadute positive anche sul piano della apertura del processo di ricerca [18]). I noti problemi di "allucinazioni" – la produzione di risposte false a livello referenziale ancorché linguisticamente accettabili – e l'inesattezza nei riferimenti bibliografici e fattuali ovviamente vanno tenuti nel dovuto conto, e impediscono (per fortuna diremmo) di delegare totalmente il processo di ricerca. Tuttavia, una accurata predisposizione dei prompt, mediante quello che

chiameremmo un *research oriented prompt engineering*, può ovviare almeno in parte a questi problemi: ci riferiamo in particolare all'uso di tecniche di *contextual prompting* [19] e di *Chain-of-Thought prompting* [20] che sono state sviluppate e testate di recente come alternative a costose complesse e non sempre efficaci fasi di fine-tuning, con risultati notevoli.

I modelli linguistici conversazionali possono dunque facilitare la costituzione di un *common ground* concettuale e metodologico tra ricercatori formati in domini differenti, prerequisito necessario per l'efficacia della ricerca interdisciplinare che caratterizza ogni settore delle Digital Humanities. E di conseguenza possono anche essere un fattore di innalzamento della produttività del lavoro scientifico<sup>4</sup>.

Assai più controverso è l'uso di LLM nella fase di interpretazione dei risultati di generalizzazione e di spiegazione causale dei risultati dell'analisi dei dati. In un certo senso presumere che un modello linguistico possa fornire autonomamente risultati in questi ambiti presuppone che esso sia dotato di una qualche capacità metacognitiva, mentre è ancora ampiamente dibattuto se essi siano, anche in misura molto limitata, qualcosa di più di cosiddetti *stochastic parrot*, pappagalli casuali [22]. Non è mia intenzione in questo contesto entrare nel dibattito filosofico circa la natura intelligente per non dire poi cosciente o come si usa dire oggi senziente di questi modelli [23], [24]. Di sicuro essi sono in grado di funzionare in un modo sorprendente rispetto alle stesse aspettative di coloro che li hanno progettati. A ben vedere molte delle critiche cui vengono soggetti sono fondamentalmente variazioni su posizioni metafisiche dualiste cartesiane, che ritenevamo oramai superate, o di naturalismo biologico radicale.

Certo è che assegnare funzioni di valutazione e spiegazione a un modello linguistico ancorché avanzato come GPT e i suoi simili richiederebbe una adeguata soluzione al problema della esplicabilità dei risultati dei modelli stessi [25], [26], problema al centro della ricerca sulla *explainable AI* e tutt'altro che risolto. In linea di principio possiamo immaginare di chiedere al modello di spiegare tutte le ragioni delle sue scelte; il fatto è che, come studiosi e ricercatori, in un contesto di indagine scientifica non possiamo assumere le motivazioni discorsive addotte dal modello stesso come giustificazione di una spiegazione causale o di una interpretazione dei risultati di un esperimento. Non è ancora chiaro, infatti, se questi modelli abbiano sostanziali capacità di astrazione (non sto sostenendo che *non* ne abbia, solo che *non sappiamo* quanta e come ne abbia). Siamo dunque di fronte al limite epistemologico di sistemi la cui evoluzione tecnica è stata così rapida da superare di gran lunga la nostra capacità di interpretazione e spiegazione teorica di questi stessi sistemi. Tuttavia, è assai ragionevole pensare che sia l'evoluzione della ricerca nell'ambito stretto del *machine learning* e dell'IA, attraverso la produzione di modelli sempre più potenti ed efficienti, sia la riflessione teorica sugli stessi fornirà risposte a molte delle domande che rimangono allo stato attuale aperte e "flottanti". Per il momento possiamo comunque essere più che soddisfatti di come questi Pappagalli Stocastici possono aiutare noi, Veri Ricercatori, nel progettare buona ricerca umanistico computazionale.

## BIBLIOGRAFIA

- [1] OpenAI. «Introducing ChatGPT». Consultato 5 marzo 2023. <https://openai.com/blog/chatgpt>
- [2] OpenAI. «GPT-4 Technical Report», 2023. <https://doi.org/10.48550/ARXIV.2303.08774>
- [3] Marche, Stephen. «The College Essay Is Dead. Nobody is prepared for how AI will transform academia». The Atlantic, 6 dicembre 2022. <https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371>
- [4] «Tools Such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use». Nature 613, fasc. 7945 (26 gennaio 2023): 612–612. <https://doi.org/10.1038/d41586-023-00191-1>
- [5] «arXiv announces new policy on ChatGPT and similar tools – arXiv.org blog». Consultato 5 marzo 2023. <https://blog.arxiv.org/2023/01/31/arxiv-announces-new-policy-on-chatgpt-and-similar-tools/?fbclid=IwAR05arV1UDPVN5bb1Bh4MchaylYxWveYfX9mslCUHCKY8-Sm4Fe5gwGmKYM>
- [6] MLA Style Center. «How Do I Cite Generative AI in MLA Style?», 17 marzo 2023. <https://style.mla.org/citing-generative-ai/>
- [7] Underwood, Ted. *Distant horizons: digital evidence and literary change*. Chicago: The University of Chicago Press, 2019.
- [8] Gavin, Michael. *Literary mathematics: quantitative theory for textual studies*. Stanford text technologies. Stanford, California: Stanford University Press, 2023.
- [9] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, e Illia Polosukhin. «Attention Is All You Need», 2017. <https://doi.org/10.48550/ARXIV.1706.03762>
- [10] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. «Language Models are Few-Shot Learners». In *Advances in Neural Information Processing Systems*, a cura di H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, e H. Lin, 33:1877–1901. Curran Associates, Inc., 2020. [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf)

<sup>4</sup> Ci sono numerose analisi sull'impatto positivo degli LLM nella produttività in ambito di sviluppo software e un recente lavoro sperimentale di due ricercatori del MIT sui loro effetti nei compiti di redazionali di professionisti di medio livello mostra come "exposure to ChatGPT increases job satisfaction and self-efficacy and heightens both concern and excitement about automation technologies" [21].

- [11] Wolfram, Stephen. «What Is ChatGPT Doing ... and Why Does It Work?» Stephen Wolfram: Official Website, 14 febbraio 2023. <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work>
- [12] Balestriero, Randall, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, et al. «A Cookbook of Self-Supervised Learning», 2023. <https://doi.org/10.48550/ARXIV.2304.12210>
- [13] Rumelhart, David E., James L. McClelland, e AU. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press, 1986.
- [14] Mikolov, Tomas, Kai Chen, Greg Corrado, e Jeffrey Dean. «Efficient Estimation of Word Representations in Vector Space», 2013. <https://doi.org/10.48550/ARXIV.1301.3781>
- [15] Sennrich, Rico, Barry Haddow, e Alexandra Birch. «Neural Machine Translation of Rare Words with Subword Units», 2015. <https://doi.org/10.48550/ARXIV.1508.07909>
- [16] Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. «Training language models to follow instructions with human feedback», 2022. <https://doi.org/10.48550/ARXIV.2203.02155>
- [17] Ciotti, Fabio. «La svolta empirico-computazionale negli studi culturali e letterari: una nuova scienza della cultura». In *Culture digitali. Intersezioni: filosofia, arti, media*. Proceedings AIUCD2022. AIUCD - Ass. per l'informatica umanistica e la cultura digitale, 2022. <https://t.co/2WzCpVb9ho>
- [18] «Open Science Workflows | FOSTER». Consultato 5 marzo 2023. <https://www.fosteropenscience.eu/foster-taxonomy/open-science-workflows>
- [19] Swamy, Sandesh, Narges Tabari, Chacha Chen, e Rashmi Gangadharaiyah. «Contextual Dynamic Prompting for Response Generation in Task-oriented Dialog Systems», 2023. <https://doi.org/10.48550/ARXIV.2301.13268>
- [20] Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, e Denny Zhou. «Chain-of-Thought Prompting Elicits Reasoning in Large Language Models», 2023. <https://doi.org/10.48550/ARXIV.2201.11903>
- [21] Noy, Shakked, e Whitney Zhang. «Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence». *SSRN Electronic Journal*, 2023. <https://doi.org/10.2139/ssrn.4375283>
- [22] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, e Shmargaret Shmitchell. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜». In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM, 2021. <https://doi.org/10.1145/3442188.3445922>
- [23] Chalmers, David J. «Could a large language model be conscious?», manuscript.
- [24] Searle, John Rogers. *Intelligenza artificiale e pensiero umano. Filosofia per un tempo nuovo*. A cura di Angela Condello. Roma: Castelvecchi, 2023.
- [25] Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, e Bin Yu. «Definitions, Methods, and Applications in Interpretable Machine Learning». *Proceedings of the National Academy of Sciences* 116, fasc. 44 (29 ottobre 2019): 22071–80. <https://doi.org/10.1073/pnas.1900654116>
- [26] Confalonieri, Roberto, Ludovik Coba, Benedikt Wagner, e Tarek R. Besold. «A Historical Perspective of Explainable Artificial Intelligence». *WIREs Data Mining and Knowledge Discovery* 11, fasc. 1 (gennaio 2021). <https://doi.org/10.1002/widm.1391>

# Un atlante digitale per i borghi delle aree interne

Vincenzo Colaprice<sup>1</sup>, Nicola Barbuti<sup>2</sup>

<sup>1</sup> Università degli Studi di Bari Aldo Moro, Italia – vincenzo.colaprice@uniba.it

<sup>2</sup> Università degli Studi di Bari Aldo Moro, Italia – nicola.barbuti@uniba.it

## ABSTRACT

L'attenzione sempre maggiore riservata ai borghi delle aree interne ha sollecitato una domanda crescente di analisi, studio e comprensione di questi luoghi e delle possibilità di rigenerazione a partire dal patrimonio culturale. Attraverso lo sviluppo di un atlante digitale dei borghi è possibile realizzare sistemi di mappatura che consentono di restituire una fotografia della reale condizione in cui versano i borghi.

## PAROLE CHIAVE

Mappature digitali; web-map; GIS, aree interne; borghi; accessibilità; patrimonio culturale

## 1. INTRODUZIONE

Nell'ultimo decennio i borghi italiani delle aree interne hanno acquisito una centralità sempre maggiore all'interno del panorama turistico e culturale italiano. Le restrizioni imposte dalla pandemia hanno favorito l'incremento dei flussi di visitatori in questi luoghi e al contempo il manifestarsi di flussi di ritorno di lavoratori e studenti fuorisede, conseguenza diretta dei nuovi bisogni emersi nel contesto pandemico come lo *smart working* e il turismo *living site* 00. La ritrovata centralità assunta dall'argomento borghi è testimoniata anche dallo sviluppo di un ampio dibattito pubblico avviato sulla scia della definizione della *Strategia Nazionale per le Aree Interne* (SNAI) nel 2013. A tal proposito, sono comparse diverse pubblicazioni 00 che hanno messo in discussione una narrazione legata ai borghi per troppo tempo appiattita sull'esaltazione delle bellezze architettoniche e naturali, senza prendere in considerazione le problematiche legate alle condizioni di vita delle comunità locali, alla scarsa accessibilità e alla carenza di infrastrutture e servizi 000. Un ulteriore passo avanti è stato compiuto recentemente dal Ministero della Cultura con la definizione di borgo quale «piccolo comune con popolazione residente fino ai 5.000 abitanti»<sup>1</sup>. Seguendo questo criterio, ben 5.537 dei 7.904 comuni italiani sono identificabili come borghi<sup>2</sup>. Le ridotte dimensioni demografiche e le fragilità che contraddistinguono paesi e piccoli comuni delle aree interne devono condurre ad un approccio differente rispetto al tema “borghi”, concependo questi luoghi come ecosistemi culturali, ovvero espressioni di un patrimonio culturale ramificato, nel quale si intrecciano patrimonio architettonico, artistico, immateriale e naturale. I borghi sono luoghi in cui la vita e l'identità della comunità locale si legano in maniera profonda al tessuto culturale, paesaggistico e produttivo del territorio 0. Alla luce di questo contesto così composito che riguarda circa un sesto della popolazione italiana e coinvolge un terzo del patrimonio culturale, nella letteratura recente sui borghi è emersa la necessità di mappare il territorio in maniera analitica per fornire elementi che costruiscano una visione comune nell'impegno di istituzioni, cittadini e comunità locali. Serve dunque una pluralità di mappe che tracci grandi «concettualizzazioni del territorio» e ponga «l'enfasi sui luoghi» a partire dal patrimonio culturale e naturale, dalla demografia e dalla rete di servizi esistenti 0. Le mappe diventano dunque un modo per orientare politiche *placed-based* e sviluppare nuove sinergie per coinvolgere gli utenti finali e le comunità locali.

## 2. L'ATLANTE DIGITALE DEI BORGHI

Nonostante le numerose pubblicazioni e progettualità degli ultimi anni legate ai borghi, non esiste ad oggi una mappatura puntuale e aggiornata dei piccoli comuni italiani che possa offrire differenti livelli di analisi. L'idea di fondo è quella di dare vita ad un atlante digitale dei borghi che possa includere una molteplicità di aspetti che trasmettano il senso della complessità di questi luoghi e il legame profondo che essi hanno con il patrimonio culturale e naturale, senza tralasciare le criticità del territorio. L'atlante digitale dei borghi può contribuire a ridefinire l'approccio a questi luoghi, troppo spesso caratterizzato da processi che puntano unicamente alla valorizzazione commerciale dei borghi turistici. Al contrario, l'alto numero di comuni italiani considerabili come borghi dovrebbe orientare la riflessione verso altri elementi che caratterizzano quotidianamente la vita delle comunità locali, le esigenze dei cittadini e le relazioni che intercorrono con il territorio circostante e i centri urbani più vicini. L'atlante può offrire una prospettiva alternativa, evidenziando tutti quegli aspetti

<sup>1</sup> A seguito di un lungo dibattito, il Ministero della Cultura ha definito “borghi” i piccoli comuni singoli o aggregati che hanno «popolazione residente fino a 5000 abitanti» (vedasi <https://cultura.gov.it/borghi>, Bando PNRR Borghi-Linea B).

<sup>2</sup> Dati ISTAT al 01/01/2023.

che possono restituire una fotografia della condizione in cui versano i borghi, mettendo in risalto dati e indicatori relativi al patrimonio culturale e naturale, all'andamento demografico, alle condizioni di accessibilità, alla presenza all'interno di reti e associazioni nazionali, la contiguità rispetto a percorsi di mobilità sostenibile. Per realizzare l'atlante è fondamentale attingere a dati pubblici e liberamente accessibili, trasformando l'atlante in un collettore di dati solitamente dispersi e non messi in correlazione tra loro. Nella realizzazione dell'atlante si è usufruito di una pluralità di fonti che rilasciano dati relativi ai vari aspetti presi in considerazione. La banca dati ISTAT occupa un posto di rilievo tra le fonti utilizzate, fornendo dati relativi agli aspetti demografici e culturali; in secondo luogo si è attinto ai dati e alle rappresentazioni vettoriali contenute nei piani paesaggistici delle regioni italiane per quanto concerne i beni naturali; in terzo luogo, dovendo censire anche il patrimonio culturale immateriale e avendo necessità di affidarsi a criteri affidabili, si è preferito visualizzare i beni culturali immateriali censiti da progetti ministeriali; un'ulteriore fonte è rappresentata dai Linked Open Data del Ministero della Cultura, attraverso i quali è possibile censire e rappresentare le istituzioni culturali presenti nei territori dei borghi; infine, attraverso l'impiego di un algoritmo sviluppato in ambiente GIS è possibile valutare il grado di accessibilità fisica ai borghi delle aree interne e le distanze dai principali nodi infrastrutturali e dai centri urbani maggiori. L'atlante si configura come un sistema di web-map interattive e multilivellari attraverso le quali è possibile visualizzare informazioni relative all'andamento demografico, all'ubicazione dei borghi, al grado di accessibilità a questi luoghi, alla presenza di infrastrutture e servizi adeguati, all'appartenenza dei borghi all'interno di reti associative di carattere nazionale, al coinvolgimento di questi comuni all'interno delle misure varate dal PNRR e infine, ma non per minore importanza, mappare la presenza del patrimonio culturale, materiale e immateriale, a partire da numerosi indicatori che possono riguardare la presenza di beni archeologici e architettonici, la prossimità di beni UNESCO e di parchi e aree protette, nonché l'insistenza di minoranze linguistiche. L'atlante digitale dei borghi muove da una riflessione più ampia condotta in relazione al concetto di accessibilità: l'interazione con i sistemi di mappatura integrati nell'atlante consentirebbe ad una vasta comunità di utenti di accedere alla conoscenza dei borghi e delle aree interne a partire informazioni relative agli aspetti demografici, culturali e naturalistici; in secondo luogo, le mappature digitali possono aiutare gli utenti a definire il grado di accessibilità ai borghi, studiandone gli itinerari, le soluzioni disponibili in termini di trasporto pubblico e mobilità sostenibile, la sicurezza dei percorsi di visita; infine, l'accessibilità di questi luoghi può essere compresa anche alla luce delle condizioni di vita, comprendendo le tipologie di servizi essenziali presenti sul territorio e le vocazioni di ciascun luogo. Stante il continuo sviluppo del progetto, l'idea è caratterizzare ulteriormente l'attività di mappatura attraverso l'inclusione di espansioni digitali che possano aumentare il livello di interattività delle mappe e consentire a molteplici comunità di utenti di accedere a svariate tipologie di informazioni e rappresentazioni dei beni culturali legati ai borghi.

### 3. LA REALIZZAZIONE

Lo sviluppo delle mappature digitali ha seguito diverse fasi. La fase iniziale ha previsto la creazione di dataset contenenti tutti i dati relativi ai comuni italiani con popolazione non superiore ai 5.000 abitanti. I dati sono stati estratti dal portale dell'Istituto Nazionale di Statistica (ISTAT) e riguardano la dimensione demografica e i beni culturali dei comuni presi in considerazione. I dati sono stati successivamente elaborati all'interno del software QGIS attraverso il quale si sono definiti gli stili di visualizzazione delle informazioni raccolte. Gli Shapefile e i file GeoJSON ottenuti dalla trasformazione vettoriale dei dati sono stati esportati e caricati all'interno dello spazio web nel quale sono state sviluppate le web-map multilivellari. Le web-map sono state realizzate in linguaggio HTML e personalizzate in CSS sebbene sia stato fondamentale il ricorso alla library Leaflet-JS che consente di personalizzare in profondità le mappe e garantire un livello di interazione elevato. All'interno delle mappe è possibile consultare le informazioni relative ai diversi luoghi attraverso la consultazione dei popup, cercare i luoghi desiderati attraverso una barra di navigazione e impostare gli itinerari necessari per comprendere l'accessibilità e le modalità di raggiungimento dei luoghi. Ogni luogo ed indicatore presente all'interno delle mappe è identificato da un segnaposto (*marker*) che consente di accedere ad un popup i cui contenuti differiscono in base alla tipologia di mappa realizzata. A questo si aggiunge la presenza di forme vettoriali che rappresentano l'estensione territoriale dei comuni o i beni paesaggistici all'interno del contesto regionale. Per conferire un esempio si veda il caso del comune di Celle San Vito, borgo con 148 abitanti, il meno popolato della Puglia. Un primo set di mappe rappresenta i dati demografici e la diffusione dei borghi. Ogni mappa regionale contiene tre livelli (Figura 1): confini amministrativi di tutti i comuni; puntinato dei borghi; poligoni della superficie dei borghi. Il puntinato permette di geolocalizzare i borghi nella regione. I segnaposto, quando cliccati, offrono dati relativi alla demografia del borgo. L'uso dei poligoni consente di comprendere la dimensione demografica dei borghi in relazione al contesto spaziale. Il numero di abitanti dei borghi è tripartito in tre fasce, contraddistinte dall'uso di colori differenti: rosso da 0 a 1.000 abitanti, giallo da 1.001 a 2.500, blu da 2.501 a 5.000. Un secondo set di mappatura riunisce indicatori differenti relativi al patrimonio culturale dei borghi messo in relazione con il patrimonio paesaggistico della regione. Il segnaposto dei borghi consente di accedere attraverso il popup alla consultazione dei dati quantitativi relativi alla presenza dei beni culturali censiti da ISTAT in ogni comune.



A questi dati, si aggiungono segnaposto che segnalano la prossimità di beni UNESCO e del patrimonio culturale immateriale censito da progetti ministeriali, quali eventi, saperi e tradizioni. Il patrimonio paesaggistico è rappresentato attraverso la visualizzazione di forme vettoriali raffiguranti parchi e aree protette riconosciute da Stato e Regione, nonché fiumi, laghi e monti al di sopra dei 1.200 metri. Gli Shapefile sono utilizzati anche per segnalare la presenza di minoranze linguistiche, elevando il grado di informazioni che la mappa dei beni culturali può offrire. Ulteriori tipologie di mappatura riguardano la sovrapposizione tra segnaposto dei borghi e itinerari dei cammini d'Italia, oltre a mappe riassuntive delle reti di promozione nelle quali i borghi sono coinvolti. Nell'ambiente GIS è stata sviluppata una *network analysis* che consente di identificare il grado di accessibilità dei borghi in relazione alle infrastrutture e ai centri principali. Attraverso una colorazione differenziata dei segnaposto è possibile valutare quali luoghi fronteggino maggiori o minori criticità in termini di accessibilità fisica e visualizzare i dati relativi alla distanza dalle reti di trasporto principali.

## BIBLIOGRAFIA

- [1] Carrera, Letizia e Barbuti, Nicola. "After the pandemic. A 4.0 model for the recovery of the Cultural tourism." *International Journal of Humanities and Social Science* 11, no. 4 (April 2021): 47-58.
- [2] Carrera, Letizia e Barbuti, Nicola. «Ripensare l'offerta turistica e culturale negli scenari COVID-19. Una sfida aperta.» In *Riflessioni intorno al viaggio turistico delle radici. Esperienze, strategie e scenari post COVID-19*, a cura di Letizia Carrera, Tullio Romita e Antonella Perri, 15-44. Bruxelles: Etqua, 2020.
- [3] Cersosimo, Domenico, e Donzelli, Carmine (a cura di). *Manifesto per Riabitare l'Italia*. Roma: Donzelli Editore, 2018.
- [4] De Rossi, Antonio (a cura di). *Riabitare l'Italia: le aree interne tra abbandoni e riconquiste*. Roma: Donzelli Editore, 2019.
- [5] Barbera, Filippo, Cersosimo, Domenico, e De Rossi, Antonio (a cura di). *Contro i borghi. Il Belpaese che dimentica i paesi*. Roma: Donzelli Editore, 2022.
- [6] Associazione Borghi Autentici d'Italia, con il contributo delle regioni Calabria, Emilia-Romagna, Piemonte, Puglia, curata da Doxa in collaborazione con Mercury. «Indagine sui visitatori attuali e potenziali dei borghi caratteristici.» 2009.
- [7] Gugliotta, Francesca. «Di cosa parliamo quando parliamo dei borghi.» *Repubblica.it*, 4 Aprile 2010.
- [8] Cersosimo, Domenico. «Luoghi.» In *Manifesto per Riabitare l'Italia*, a cura di Domenico Cersosimo e Carmine Donzelli, 145. Roma: Donzelli Editore, 2018.
- [9] Carrosio, Giovanni e Faccini, Andrea. «Le mappe della cittadinanza nelle aree interne.» In *Riabitare l'Italia*, a cura di Antonio De Rossi, 53. Roma: Donzelli Editore, 2019.

# To what extent are lemmatisation and annotation relevant for deep learning assignments and textual motifs detection?

## The case-study of Peter Damian's letters (11<sup>th</sup> century)

Valérie Thon<sup>1</sup>, Laurent Vanni<sup>2</sup>, Dominique Longrée<sup>3</sup>

<sup>1</sup> Université Paris Cité and Université de Liège, France and Belgium – valerie.thon@u-paris.fr

<sup>2</sup> Université de Nice – Sophia Antipolis, France – laurent.vanni@unice.fr

<sup>3</sup> Université de Liège, Belgium – dominique.longree@uliege.be

### ABSTRACT

This paper wishes to explore to what extent lemmatisation and morphosyntactic annotation are important for deep learning predictions and textual motif detection. A broader research on the style of Peter Damian's letters (11<sup>th</sup> century) was the occasion to explore this question. After having trained two deep learning models on a selection of 12 classical authors using the *Hyperdeep* platform, one on lexical forms alone and the other on lemmatised and annotated texts, we introduced to them the medieval letters of Peter Damian in order to not only examine which authors are deemed to be stylistically close to Peter according to both models, but also to compare whether the results are similar and whether the same linguistic structures receive a high activation rate. The results suggest that a dialogue between both methods could be an interesting path to explore in the search for textual motifs, as the first “lexical” model may indicate rough outlines of these motifs, whereas the second model can offer concrete examples and/or variants of the first motifs identified.

### KEYWORDS

Deep learning; textual motif; lemmatisation; annotation; Peter Damian

## 1. INTRODUCTION AND METHODOLOGY

In this paper, we wish to explore the efficiency of trained deep learning models in the identification of textual motifs (Longrée et al., 2008), analysing how their results might differ when confronted firstly with non-lemmatised texts (having only at their disposal the lexical forms), versus lemmatised and morphosyntactically annotated works. Does the first category already offer sufficiently reliable results, or should we consider a dialogue between the two methods? This question lies more broadly within the framework of V. Thon's thesis project, which is focused on a stylistic study of the epistolary corpus of Peter Damian (1007-1072/73), a reforming hermit of the Central Middle Ages. For the purpose of this stylistic research, 20 letters concerning the “vices” Peter perceived within the Church of his time will be lemmatised and morphosyntactically annotated using the LEI software (“LASLA Encoding Initiative”), developed by the Laboratoire d'Analyse Statistique des Langues Anciennes (“LASLA”) of the University of Liège. LEI proposes a semi-automatic procedure which provides for each individual form of a text in treatment one or more possible analyses (lemma and morphosyntactic annotation, where each subordinate sentence also receives a specific code according to its subordinating conjunction); these are then subject to a selection and a systematic verification by a confirmed philologist, who completes and/or corrects the possibilities provided. The results will later be statistically explored with the *Hyperbase Web* platform, developed by Étienne Brunet and Laurent Vanni (UMR Bases, Corpus, Langage, Université Nice Sophia Antipolis).

In order to examine how differently trained deep learning models may assist us in the identification of textual motifs, certainly useful in the context of this broader stylistic study, we will rely here on the tools provided by the same *Hyperbase* platform and confront the epistolary corpus of Peter Damian (1007-1072/73) to a collection of 12 classical authors whose life and works range from the 3<sup>rd</sup> century BC to the 1<sup>st</sup> century AD. Why this exact selection of 12 classical authors? Our exploration requires most of all a rich and varied corpus of Latin texts that are also available in a lemmatised and morphosyntactically annotated format; since our 20 letters of Peter Damian have been subjected to a precise lexical and morphosyntactic analysis by means of the semi-automatic LEI procedure, it would also be preferable that the labelling of our comparison corpus has been carried out with the same method. The LASLA, having created a large database of digitised, lemmatised and annotated Latin and Greek texts for the purposes of linguistic and literary studies, offers such a corpus. As can be seen from the selection below, however, the LASLA has mostly focused its attention on classical

Antiquity, even though they have been expanding their field of expertise in recent years to late antique and even medieval texts (for example, Peter Damian himself). Adding new texts to this corpus would require time and, although tools for automatic lemmatisation and labelling exist, such as the Collatinus-LASLA project or *MBT (Memory-Based Tagger)*, *TnT (Trigrams'n'Tags)* and *TreeTagger*, these tools often necessitate an important manual verification or don't yet reach sufficiently high levels of precision (Verkerk et al., 2020; Longrée and Poudat, 2010). Taking into account not only the available LASLA-texts, but also chronology, genre and style, we arrived at the following selection: Plautus (comedies: *Amphitruo*, *Asinaria*, *Aulularia*, *Bacchides*, *Captivi*, *Casina*, *Curculio*, *Epidicus*), Cornelius Nepos (history: *Vitae*), Cicero (dialogues: *De Officiis I-II-III*, *Laelius vel de amicitia*, *Cato maior de senectute*; speeches: *Pro A. Caecina*, *Pro M. Fonteio*, *in L. Sergium Catilinam orations I-II-III-IV*, *Pro lege Manilia*, *Pro A. Cluentio Habito*), Cesar (history: *Bellum Civile*), Sallust (history: *de Catilinae coniuratione*, *Bellum Jugurthinum*), Livy (history: *Ab Urbe Condita*), Seneca (letters: *Epistulae*; philosophical treatises: *de providentia*, *de brevitae vitae*, *de clementia*, *de otio*, *de tranquillitate animi*, *de constantia sapientis*, *de vita beata*), Quintus Curtius (history: *Historiae Alexandri magni*), Petronius (fictional "novel": *Satyricon*), Tacitus (history: *Historiae*), Pliny (letters: *Epistulae*), Suetonius (history: *Vitae Caesarum*). Plautus is chronologically quite removed from the other authors, but we still wished to include him in our selection for the nature of his Latin, which would have been quite close to spoken Vulgar Latin, both in vocabulary and grammar. Peter's own letters, always intended to be read aloud, might therefore share some subtle morphosyntactic characteristics with Plautus' work. Any poetic text was excluded, mainly because their freer linguistic structure risked altering the results of deep learning. The spelling of the non-lemmatised text files was harmonised, as well as their punctuation and layout (weak punctuation was removed and all strong punctuation was homogenised in the form of full stop punctuation marks).

To carry out the double confrontation between our classical comparison corpus and Peter Damian's letters, we used *Hyperdeep*, a deep learning model integrated in *Hyperbase Web* and trained on a classification task (authorship attribution in our case). The architecture is based on multi-channel convolutional neural networks and allows both prediction of new text and extraction of features responsible for the model's decision (Vanni et al., 2023). Using the *Hyperdeep* platform, we trained two different models on our group of classical authors: the first trained model only has access to the lexical forms of their included texts, whereas the second model can perform more in-depth stylistic analyses and classification tasks, based not only on the lexical forms, but also on the lemma's and the morphosyntactic labelling provided (following LEI procedure). Once both models were trained (with a precision rate of 98.05% and 97.91% respectively), we introduced them to 12 of Peter Damian's letters in order to examine which classical authors were identified as hypothetical literary models and, more importantly, what textual motifs this identification was based on. In the case of the second model, trained on lemmatised and labelled classical texts, the 12 letters were also introduced in lemmatised format (the lemmatisation of the remaining 8 letters will soon be completed). Of course, there are also other stylometric methods that might help us explore our research question, such as *Stylometry with R* ("*Stylo*"). *Stylo* can be used for a number of different purposes, such as the linguistic and stylistic analysis of large textual corpora, the study of an author's personal writing style and authorship study in general, but the *Stylo* package is also capable of classification tasks (such as author attribution). The program also allows users to upload their own annotated text files, but it seems that the LASLA lemmatisation files are not yet completely compatible with this program. For this reason, we have chosen to focus our research solely on *Hyperdeep*'s deep learning models, which are able to exploit the LASLA files and their high precision semi-automatic lemmatisation with morphosyntactic labelling, but it would certainly be useful to explore other stylometric methods in future research.

## 2. A CONFRONTATION BETWEEN TWO DEEP LEARNING MODELS

### 2.1. Model 1 (lexical forms)

First, 12 letters of Peter Damian were introduced one by one to the model trained on the non-lemmatised texts of the classical corpus, and it was giving the task of assigning these letters to one or more of the authors of this group. In order to check the stability of the results on a larger population, we also proposed to this model almost all of Peter's letters in a single file (*Patrologia Latina*; 107 letters were known at the time of this edition); we should note that we were unable to do this stability-check for the second model since not all of Peter's letters are lemmatised. According to deep learning predictions, the individual 12 letters and the large corpus of the combined letters are all stylistically close to Pliny, very often identified first, and to Seneca; the recognition rates for the other authors in our "classical" corpus almost never rise above 10%. By way of illustration, we show below the general profile generated by *Hyperdeep* for the almost-entirety of Peter's epistolary corpus in one file, compared to the classical authors (Figure 1). Pliny and Seneca are recognized at 39% and 34% respectively; the next author in the list is Cicero, but his recognition rate is only situated at 7.1%:

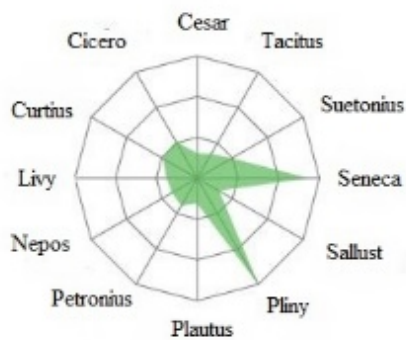


Figure 1. Recognition rates of Peter's epistolary corpus.

Within the classical corpus, Peter's work is clearly associated with that of the two other letter-writers: Pliny and Seneca. Their hypothetical identification seems to rest primarily on a shared vocabulary. For Pliny, for example, many connectors used at the beginning of a sentence are highlighted by the first *Hyperdeep* model, such as *unde* (“wherefore”), *nam* (“because”), *ut* (“so that”), *sicut* (“like”), *ergo* (“therefore”), *et* (“and/also”); some pronouns, especially of the first person, also tend to light up (*me*, *mi*, *mihi*), just like the vocative *domine* (“lord”), frequent with Pliny but used in a religious way with Peter, and different forms of the verb *dicere* (“to say”). Let us note that all these recognized forms are also characteristic of the epistolary genre in general. Connectors seem to have less weight in the recognition of Seneca; only *ergo* (“therefore”) and *quia* (“because”) occur regularly in the passages highlighted. There is, on the other hand, a whole lexicon of a rather ‘philosophical’ nature which lights up for him in Peter’s epistolary corpus (who nevertheless uses it in a religious sense): *mundus* (“world”), *persona* (“person”), *deus* (“god”), *gradus* (“degree”), *homo* (“man”), *sapiens* (“(the) wise”). We have also found several times among the words recognized for Seneca the verbal form *ait* (“he/she says”) as well as a whole series of demonstrative pronouns (*iste*, *ille*, *hic*), whose presence can be linked again to the dialogical character of the epistolary genre itself. When the latter occur, they are often preceded by a coordinating conjunction, such as *et* or, in one case, *autem*: *et iste quomodo vivet*, *et illud psalmistae*, *et illa signanter*, *et illud Ysaie*, *et illum blasphemantes*, *autem iste illa*. In some cases, the pronoun is surrounded by two verbs: *possunt ista congruere*, *facientibus ista consentiunt*, *facit errare ista perire* – as an example, we show below (Figure 2) the activation rate of the formula *possunt ista congruere* in *Letter 162*, written by Peter Damian around 1069-1072 and addressed to the archpriest and papal chancellor Peter, who is asked to support our hermit in his battle against clerical marriage and domestic partnership.

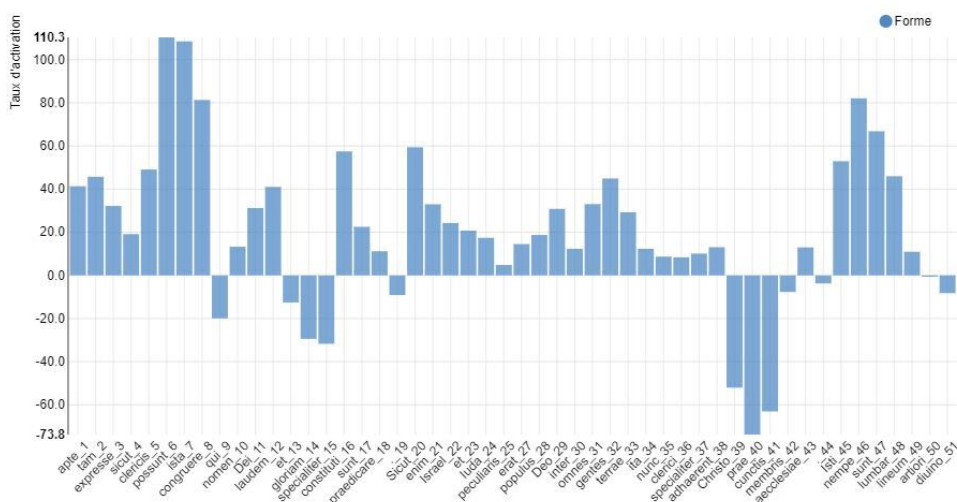


Figure 2. Recognition rate for “*possunt ista congruere*” (Letter 162).

A search in *Hyperbase's* “LASLA” database, containing almost all of the Latin texts morphosyntactically labelled by this same laboratory, tells us that some of these verbs are already slightly overused by Seneca in his work (*posse*, *facere*, *consentire*, *errare*, *perire*), but that the structure “verb – *ista* – verb” in particular is also characteristic of his style. The “LASLADamianus” database, created by us as a means to compare the already lemmatised letters of Peter with the ancient authors, confirms this first observation and also reveals that this same syntactic combination is not statistically specific of

Peter’s own style. Since previous work has highlighted the possible sensitivity of deep learning to morphosyntax (Vanni et al., 2018a and 2018b; Thon et al., 2022), it is possible that *Hyperdeep* identified in Peter’s letters a sequence “verb – demonstrative pronoun – verb” which echoes the style of Seneca. It is difficult to say whether this is a conscious borrowing on the part of our hermit. The editor of his letters, Kurt Reindel, does not cite the works of Seneca at any point among the sources used by Peter; similarly, the 12<sup>th</sup> century catalogue of the library of Fonte Avellana, the hermitage in northern Italy where Peter was not only monk but also prior, does not mention manuscripts containing texts belonging to Seneca; secondary literature also suggests that the dissemination of his work in general would have been very limited until the last decade of the 11<sup>th</sup> century and the beginning of the 12<sup>th</sup> century (Mayer, 2015, p. 277-278). The apocryphal correspondence between Seneca and St. Paul, on the other hand, was quite popular during the Middle Ages, appearing in multiple manuscripts from the 9<sup>th</sup> to the 12<sup>th</sup> century and onwards (Fürst, 2014, p. 213) – but again, no direct link between Peter’s work and this collection can be found, and these ‘false’ letters of Seneca were not included in our comparison corpus. In any case, whether Peter Damian consciously borrowed the “verb – *ista* – verb” sequence from Seneca or not, deep learning was able to reveal its presence.

## 2.2. Model 2 (lemmatised and annotated)

In order to further explore these first results based solely on lexical forms, which are already quite promising, we have recreated our primary corpus of 12 ancient authors, but this time including only their works lemmatised and labelled with a complete morphosyntactic analysis by the LASLA by means of their LEI interface. After having trained a new deep learning model on this second, labelled, corpus representing Antiquity, we reintroduced the same 12 letters of Peter Damian – also lemmatised – to this collective of authors in order to verify 1) whether *Hyperdeep*’s predictions remain stable when passing from predictions based on lexical forms alone to labelled and annotated texts, and 2) whether the enriched training corpus (lexical forms, part of speech and lemma) would allow us to refine our first results.

In general, the recognition scores of Peter’s letters correspond for the most part to what *Hyperdeep* had predicted based on lexical forms alone: in almost all cases, and with considerably high percentages, Pliny and Seneca remain the first authors to be identified as hypothetical models. Let us go back to the example of *Letter 162*, used in the previous section. When examined by the first model trained, it was considered stylistically close to Pliny (with a recognition rate of 38%), followed closely by Seneca (35%) and, in third place, to Quintus Curtius at only 7.1%. When enriched with morphosyntactic information, *Hyperdeep* still associates the same letter with Seneca’s style (at a remarkable 53% recognition rate) and, to a lesser extent, with Pliny (14%), followed closely by Quintus Curtius (11%). Even though the exact percentages differ, our original trio has remained the same, suggesting that, even when working on forms alone, *Hyperdeep* already has a significant precision in its predictions.

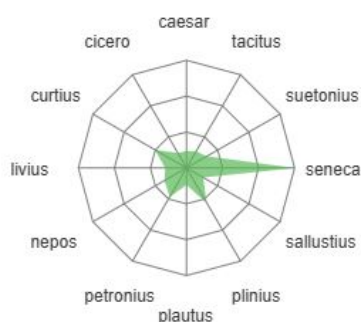


Figure 3. Recognition rate (Letter 162).

Even more interesting is the fact that, whereas deep learning based on lexical forms seemed to identify the sequence “verb – *ista* – verb” as characteristic of Seneca’s writing, the predictions dependent on the lemmatised LASLA-files allow us to push further this result by highlighting other more complex structures appearing to be variants of the same sequence. Among the key passages identified for Seneca in *Letter 162*, for example, we find the two following sentences: *ait dominus ad eum: surge, vade* and *ait omnis homo qui audit verba mea*. By means of illustration, we show here the activation rates for the first example:

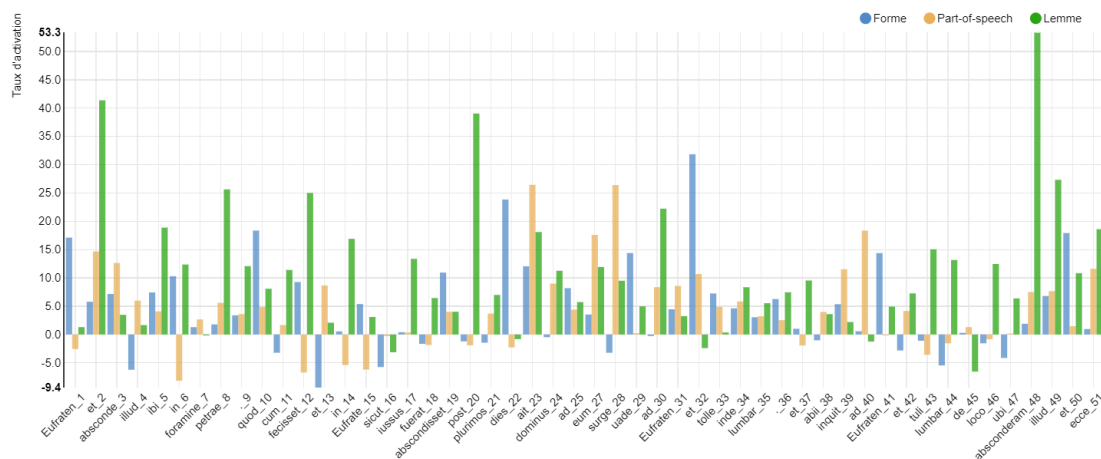


Figure 4. Activation rate for the sequence “ait dominus: ad eum vade, surge”.

Even though they appear to be different from each other at first glance, they are in essence quite similar to the sequence “verb – demonstrative or relative pronoun – verb” suggested by our previous deep learning model. In fact, both examples start out with a verb (in this case: *ait* (“he/she says”), already specific to Seneca as a lexical form alone), followed closely thereafter by a demonstrative or relative pronoun (*eum/qui*), immediately followed again by another verb (*vade/audit*), who both primarily light up on the grammatical plane. Other letters of Peter’s test-corpus present different realisations of this same textual motif as well. *Letter 61*, for instance, written around 1059 and asking pope Nicolas II to take action against bishops who are married or living in concubinage, is also estimated to be close to Seneca (59%), to Quintus Curtius (12%) and to Pliny (8.7%) according to our second model. Among the words and morphosyntactic combinations identified for Seneca in particular, the underlined words of the following sentence were activated: *apostolo testante qui ait qui adheret meretrici unum corpus efficitur*, which gives us again the “verb – relative pronoun – verb” sequence with *ait qui adheret*, where *ait* introduces once more the rest of the motif (we might therefore consider the pattern “*ait* – pronoun – verb” to be a specific variant of the more general sequence identified). Another example offered by *Letter 61* is not introduced by *ait* specifically: *Facti siquidem culpam habet qui quod potest negligit emendare*. The sequence, however, is still present; the two verbs *habet* (“he has”) and *negligit* (“he neglects”) are activated for Seneca, not only on a grammatical level, but also because of their lemma’s. They encircle the relative pronoun *quod* (“which”), which is mainly recognised for its grammatical quality as a pronoun. Using the proven methods of the more traditional TDA (Textual Data Analysis), we were also able to verify that the textual motif “verb – demonstrative or relative pronoun – verb” is still highly specific to Seneca’s oeuvre even if other elements are inserted between the verbs and the pronoun. It should however be noted that three of our four examples are of a biblical nature (see most notably Jer 13, 6-10; Mt 7, 24; 1 Cor 6, 16). Peter Damian was of course an important ecclesiastical figure who was intimately familiar with Holy Scripture, and his personal writing style was most probably influenced by this as well. A possible explanation might be that the grammatical pattern “verb – pronoun – verb” often results in quite a colloquial phrasing (“he who hears”, “he who adheres”, “he who can do this”), such as we have seen above: *homo qui audit; qui adheret; qui quod potest*, and that such language is of course similar to the writing style of the Bible, and also leaves its traces in Peter’s own Latin. It might therefore certainly be interesting to explore, in future research, the possible similarities between Seneca’s epistolary corpus and the biblical (mostly New Testament) writing style.

### 3. CONCLUSION

These results, however provisional they may be, indicate first of all the remarkable precision and sensitivities of deep learning predictions when working with lexical forms alone; according to our first trained model on ancient texts, indeed, Pliny and Seneca were the authors deemed most similar to Peter Damian’s letters. The hypothetical identification in itself makes sense, especially since all three of them have practised and are represented by the epistolary genre. The passages brought forward by the activation rates suggested that the identification relied mostly on the individual or combined specificities of lexical forms, but that *Hyperdeep* in this case may be able to sense some morphosyntactic information as well (such as verbs, for example). The second model allowed us to refine these results. Trained on the same corpus, but this time lemmatised and annotated, it not only produced a similar prediction (Seneca and Pliny), but offered far more specific information on the activation rates. More precisely, it allowed us to identify more complex variants of the previously suggested motif “verb – demonstrative or relative pronoun – verb”, three of them starting with *ait* (“he/she

says”), already characteristic in itself as a lexical form for Seneca. The confrontation with the second trained model therefore shows the interest of going past forms alone to focus on the lemma’s and on finer morphosyntactic annotations as well, and suggests that a dialogue between the two methods may be an interesting path to explore in the search for textual motifs: a model trained on forms alone in an effort to find rough outlines of these motifs, and a second model trained on three linguistic layers that would offer more concrete examples and/or variants of the motifs first identified.

## REFERENCES

- [1] Brunet, Étienne and Vanni, Laurent. «Deep learning et authentification des textes». *Texte ! Textes et cultures* 24.1 (2019): 1-34.
- [2] Eder, Maciej, Rybicki, Jan and Kestemont, Mike. «Stylometry with R: A Package for Computational Text Analysis». *The R Journal* 8.1 (2016): 107-122.
- [3] Fürst, Alfons. «Epistulae Senecae ad Paulum et Pauli ad Senecam». In *Brill’s Companion to Seneca. Philosopher and Dramatist*, edited by Andreas Heil and Gregor Damschen, 213-214. Leiden-Boston: Brill, 2014.
- [4] Longrée, Dominique and Poudat, Céline. «New Ways of Lemmatizing and Tagging Classical and post-Classical Latin: the LATLEM project of the LASLA». In *Proceedings of the 15th International Colloquium on Latin Linguistics*, edited by Peter Anreiter and Manfred Kienpointner, 683-694. Innsbruck, 2010.
- [5] Longrée, Dominique, Xuan, Luong and Mellet, Sylvie. «Les motifs : un outil pour la caractérisation topologique des textes». In *Actes des JADT 2008, 9èmes Journées internationales d’Analyse statistique des Données Textuelles*, edited by Serge Heiden and Bénédicte Pincemin, 733-744. Lyon: Presses Universitaires de Lyon, 2008.
- [6] Martini, Paola Supino. «L’inventario del secolo XII della biblioteca di Santa Croce di Fonte Avellana». In *Studi sulle società e le culture del Medioevo per Girolamo Arnaldi*, edited by Ludovico Gatto and Paola Supino Martini, 629-642. Firenze: All’Insegna del Giglio, 2002.
- [7] Mayer, Ronald. «Seneca Redivivus: Seneca in the Medieval and Renaissance World». In *The Cambridge Companion to Seneca*, edited by Shadi Bartsch and Alessandro Schiesaro, 277-288. Cambridge: Cambridge University Press, 2015.
- [8] Reindel, Kurt (éd). *Die Briefe des Petrus Damiani. Teil 1 – 4. Die Briefe der deutschen Kaiserzeit (MGH)*. München: Monumenta Germaniae Historica, 1983-1993.
- [9] Thon, Valérie, Vanni, Laurent and Longrée, Dominique. «Le deep learning auxiliaire de l’ADT dans le choix de textes à étiqueter en vue d’un corpus de comparaison : à propos de l’étude stylistique des lettres de Pierre Damien». In *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data (2 vols.)*, edited by Michelangelo Misuraca, Germana Scepi and Maria Spano, 834-841. Naples-Cosenza: VADISTAT Press/Edizioni Erranti, 2022.
- [10] Vanni, Laurent, Corneli, Marco, Longrée, Dominique, Mayaffre, Damon and Precioso, Frédéric. «Hyperdeep : deep learning descriptive pour l’analyse de données textuelles». *Lexicometrica* (2020): 1-12.
- [11] Vanni, Laurent, Corneli, Marco, Mayaffre, Damon and Precioso, Frédéric. «From text saliency to linguistic objects: learning linguistic interpretable markers with a multi-channels convolutional architecture». *Corpus* 24 (2023). <https://doi.org/10.4000/corpus.7667>
- [12] Vanni, Laurent, Ducoffre, MMélanie, Mayaffre, Damon, Precioso, Frédéric, Longrée, Dominique, Elango, Veeresh, Buitrago, Nazly Santos, Gonzalez, Juan, Galdo, Luis and Aguilar, Carlos. «Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis». In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 548-557. Melbourne, 2018.
- [13] Vanni, Laurent, Mayaffre, Damon and Longrée, Dominique. «ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables». In *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, edited by Domenica Fioredestella Iezzi, Livia Celardo and Michelangelo Misuraca, 459-466. Rome: UniversItalia, 2018.
- [14] Verkerk, Philippe, Ouvrard, Yves, Fantoli, Margherita and Longrée, Dominique. «L.A.S.L.A. and Collatinus: a convergence in lexica». In *Studi e saggi linguistici*, edited by Laura Tesconi, 1-26. Pisa: Edizioni ETS, 2020.



### **3. Intelligenza artificiale e modelli applicati ai beni culturali**

# ***Falso movimento* (Moretti 2022): Memories and Translations**

Claire Clivaz<sup>1</sup>

<sup>1</sup> DH+, SIB Swiss Institute of Bioinformatics, Switzerland – [claire.clivaz@sib.swiss](mailto:claire.clivaz@sib.swiss)

## **ABSTRACT\***

The last monograph of Franco Moretti – *Falso movimento* – was published more than a year ago, in January 2022, but has so far only been translated in German, and not in English. Regarding the international success of *La letteratura vista da lontano* (2005), this fact is surprising. Several reasons could explain the delay in publishing an English translation; this paper will highlight a point that could make the reception of this new monograph uncomfortable in English language scholarship: it clearly expresses doubts about how the distant reading has shaped the humanities. These doubts can only surprise a large part of English language scholarship: it majorly considers *La letteratura vista da lontano* as a pertinent and iconoclast history of literature. Paying attention to Moretti’s heritage and reception in diverse languages highlights the importance of *Falso movimento*: it should be considered as challenging to scholars both within and outside of the digital humanities.

## **KEYWORDS**

Moretti; Humanities; Interpretation; Distant Reading; *Falso Movimento*

## **1. INTRODUCTION**

The last monograph of Franco Moretti – *Falso movimento* – was published more than a year ago, in January 2022, but has so far only been translated in German, and not in English (as of early May 2023, to my knowledge). Regarding the international success of *La letteratura vista da lontano* (2005) in the digital humanities, this fact is surprising.<sup>1</sup> It is even more striking that *La letteratura* was published in its original Italian language and in its English translation in the same year. It is therefore appropriate to present an analysis of this multilingual case at the AIUCD conference 2023 from a non-Italian perspective but based on the original Italian work of Moretti. Several reasons can explain a delay in publishing an English translation, from practical ones to complex possibilities. This paper will highlight an issue that could potentially make the reception of this new monograph uncomfortable in English language scholarship: it clearly expresses doubts about how this distant reading has shaped the humanities.<sup>2</sup> These doubts can only surprise a large part of English language scholarship: it majorly considers *La letteratura vista da lontano* as a pertinent and iconoclast history of literature.

The general tone used by Moretti in *Falso movimento* illustrates lucid and cold memories of “il camino percorso” – and describes how far we have come with the distant reading: “The point is not to feel nostalgia, and even less to recriminate, but to understand.”<sup>3</sup> Even if nostalgia is not on Moretti’s agenda, his evaluation is quite reserved. He asserts that the digital humanities have lost the form and the social dimension of the literature, “reducing it to a pale reflection.”<sup>4</sup> Such a perspective enters into direct conflict with the more pragmatic English language perspective, as *Falso movimento* posits in the conclusion: if many changes took place in the digital humanities, one did not see *the* transformation “that seemed possible at a certain moment. It is maybe because the digital humanities escaped the confrontation with the big aesthetic and scientific culture *del Novecento*, by preferring the narrower perimeter of the recent American criticism.”<sup>5</sup> The last

---

\* Thanks are due to Andrea Stevens for her English proofreading of this article. I also thank the anonymous reviewers of this article for their suggestions.

<sup>1</sup> The 2023 article by Primorac et al. illustrates well the case: the work of Franco Moretti is at the center of this article entitled “Distant Reading Two Decades on: Reflections on the Digital Turn in the Study of Literature,” but *Falso movimento* is not discussed, and not even quoted in the bibliography.

<sup>2</sup> Several authors have of course discussed the distant reading concept and effects, a discussion beyond the scope of this short article. See e.g. Da, “The Computational Case against Computational Literary Studies” (2019); Piper, “Do We Know What We Are Doing?” (2020); Ciotti, “Distant Reading in Literary Studies: A Methodology in Quest of Theory” (2021); Primorac et al., “Distant Reading” (2023).

<sup>3</sup> Moretti, *Falso*, 11: “Non si tratta di avere nostalgia, e men che meno di recriminare. Si tratta di capire.” The English translations are mine.

<sup>4</sup> Moretti, *Falso*, 11: “Si perde la forma, si perde la dimensione sociale della letteratura, e la si riduce a uno scialbo riflesso.”

<sup>5</sup> Moretti, *Falso*, 149: “Ma dei singoli cambiamenti non sono *il* cambiamento che era un tempo sembrato possibile. Forse, è perché le *digital humanities* hanno fuggito il confronto con la grande cultura estetica e scientifica del Novecento, preferendo l’assai più angusto perimetro della critica americana recente.”

sentence could be accompanied by the Italian song *Tu vuò fà l'americano!*<sup>6</sup> Let's recognize it: the recent monograph of Moretti confronts the "American" perception of the digital humanities, and this requires further analysis.

## 2. LA LETTERATURA (2005) AND ITS TRANSLATIONS

Published in 2005 in both its original Italian language and its English translation, *La letteratura vista da lontano* has become a kind of textbook of literary reading in English language scholarship, as the subtitle suggests: *Graphs, Maps, Trees: Abstract Models for a Literary History*. It is interesting to note that the French translation, published three years later (2008), reformulates the English subtitle in its own way: *Graphes, cartes et arbres: Modèles abstraits pour une autre histoire de la littérature*. There is obviously nuance between presenting "a literary history," or "une autre histoire de la littérature," ("another history of the literature"). The English subtitle suggests that a key has been found to present the literary history: it illustrates the hope that Moretti said was lost in 2022, the transformation "that seemed possible at a certain moment."<sup>7</sup> This victorious perception of the distant reading is a major way in which *La letteratura* has presented itself throughout the English translation, and has also influenced the French reception, as it is clear in the 2022 French collected essays *La fabrique numérique des corpus en sciences humaines et sociales*. In it, the heritage of Moretti is presented by Juliette Mayer as a major and uncontested methodology: "[Working] on big scale corpora implies methodologies and tools that give a considerable space to the 'distant reading' – methodologies and computing tools that allow to reveal recurrent patterns, schemes, tendencies in big textual corpora – transformed in 'data' analyzable, quantifiable, and searchable by algorithms (Moretti 2005)."<sup>8</sup> The same vision is shared by Bénédicte Terrisse and Werner Wögerbauer, who mention in 2022 the publication of *Falso movimento* in a footnote, but without integrating it in their argumentation.<sup>9</sup>

But it should be highlighted that Moretti has maintained a prominent role to the close reading in *Graphs, Maps and Trees*, if it is read not only from a pragmatic point of view. I have always considered this 2005 book not as a textbook of the distant reading, but as a prudent essay, keeping a strong concern for interpretative core skills needed by humanists. It should even be noted that the third part – *Trees* – focuses on the "micro-level of stylistic mutations"<sup>10</sup>: there are definitively close-reading elements in this book. Moreover, Moretti describes, as a "major issue of its own right," the fact that "the models I have presented share a clear preference for explanation over interpretation; or perhaps, better, for the explanation of general structures over the interpretation of individual texts."<sup>11</sup> If, in *Falso movimento*, the questions of the interpretation and individual texts come back to center stage (see Section 3), they were already present in *La letteratura*. With lucid perception, Moretti concludes *Graphs* by stating that he presented "a materialist conception of the form," a conception duly rooted in a sociological analysis that recognizes the form as a social force.<sup>12</sup> A balanced view is promoted here, including an exhaustive perception of the humanities' core skills. Eva Geulen's comments in *Falsche Bewegung*, point out the unavoidable but mysterious relationship between form and history.<sup>13</sup> But such an evaluation could also describe *Graphs, Maps and Trees*.

Considering these remarks, one can start to perceive that *Falso movimento* is closer to *La letteratura* than expected. There is no need for reassurance considering that Moretti "did not lose the faith" – according to Demetrio Marra –<sup>14</sup> or that there is "hope in a new start" – according to Danilo Soscia.<sup>15</sup> Such perspectives miss the point: *La letteratura vista da lontano* was already highlighting the possible holes and issues of the quantitative analysis. For seventeen years, Moretti has been an actor and observer of the developments of this reading from a distance. In 2022, he simply published the next portion of his thoughts, focusing on the missing points referenced in 2005. Eva Geulen considers, in a quite dramatic way, that *Falso movimento* "not only counts the loses, but also recalculates the costs [*eine Gegenrechnung aufmachen*]."<sup>16</sup> Such an opinion overlooks that the main elements of Moretti's 2022 analysis already existed, at least *in nuce*, in *Graphs, Maps and*

---

6 This song was written by Renato Carosone in 1956; [https://en.wikipedia.org/wiki/Tu\\_vuò\\_fà\\_l'27americano](https://en.wikipedia.org/wiki/Tu_vuò_fà_l'27americano)

7 See footnote 6 above.

8 de Maeyer, "Travailler les corpus de presse numérique, un travail d'artisan?", §4.

9 Terrisse and Wögerbauer, "La carte et la frise : des métaphores aux diagrammes", §9 : "L'un des représentants les plus éminents de leur utilisation pour figurer des phénomènes de l'histoire littéraire est Franco Moretti", with a reference to *Falso movimento* in footnote 33.

10 Moretti, *Graphs*, 91.

11 Moretti, *Graphs*, 91.

12 Moretti, *Graphs*, 92.

13 Geulen, "DISTANT READING UP CLOSE: Moretti zieht Bilanz", 5: "'Es gibt keine Formen ohne Geschichte' (Falsche Bewegung, 142). Und es gibt auch keine Geschichte ohne Formen. Aber wie Formen und Geschichte aufeinander bezogen werden können, bleibt ein Rätsel".

14 Marra, "Sul Falso movimento di Franco Moretti (e cioè sull'intera collana, se vogliamo)", online.

15 Soscia, "La svolta quantitativa spiegata a mio figlio", online.

16 Geulen, "DISTANT READING", 3.

*Trees*. I would say that “la facture est salée” – the bill is heavy – only for those who thought they had finally arrived in a wealthy humanist paradise, digital one.

### 3. FALSO MOVIMENTO: MEMORIES OF SEVERAL PASTS

The last remarks of Section 2 help one to understand why it is not a surprise to read Moretti, in 2022, pleading for attention to be given to the individual texts. He comments on the Ngram patterns in this way: “The text was at the center of the study of the literature: in this case, it is a little dot. It has been reduced to a little dot”; we arrived at “a story of the literature without texts. Without texts – in the sense that they are too many, naturally, and then, they cannot be studied in their individuality anymore.”<sup>17</sup> Moretti starts *Falso movimento* from the arrival point of *La letteratura*: the loss of attention to individual texts and to interpretation, as discussed in Section 2. The image of a text reduced to a “little dot” – *un puntino* – is surely a metaphor that will remain as illustration of this book, since it expresses the present feelings of a lot of digital and/or humanist scholars. It should be noted that Moretti relates the power of this form influencing the world of interpretation, instead of the quantitative explanation. The quantitative analysis can move only “tra forma e forma” and considers the form as a finished product, measurable with a calculating mindset and applicable to many kinds of relationship.<sup>18</sup> That’s surely one of the most provocative points of the book and merits in-depth testing. For the author, it is not a question, but a conviction that leads him to a strong affirmation: “Diciamo la verità: la storia della cultura merita di meglio”, “we say the truth: the history of the culture deserves better.”<sup>19</sup> For Moretti, *l’heure est grave*: loss of strong explanations and ambitious concepts; this is the “reality of the digital humanities.”<sup>20</sup> At this point, it becomes clear that *Falso movimento*’s English translation and reception will be a challenge for scholars and teachers.

Beyond the quite short retention of quantitative literary research, the title of the book refers also to a deep European literary history. As explained by Moretti himself in a 2022 blog article, the title *Falso movimento* is inspired by a German movie by Wim Wenders, *Falsche Bewegung* (1975), inspired itself by a novel of Goethe, *Wilhelm Meister’s Apprenticeship* (1795–1796). This cultural background explains why the book has already been translated in German: *Falsche Bewegung. Die digitale Wende in den Literatur- und Kulturwissenschaften* (2022). The subtitle – “the digital *Wende* in the literary and cultural humanities” – makes the reference to Wenders’ movie stronger. The German translation definitively puts its own touch on this subtitle, since *die Wende* also means the “turn-around,” a peaceful historical and political process in the former German Democratic Republic, in the years 1989 and 1990, that led to the reunification of Germany.<sup>21</sup> Is there a way to maintain a positive perception of the digital turn, despite *Falsche Bewegung*? I will let German readers confirm what kind of *Wende* they perceive in the digital turn. But such a reappropriation in another language and culture in any case demonstrates the strong impact of *Falso movimento*. It highlights the central role of Goethe and his work in the current discussions about distant reading.<sup>22</sup> The – momentaneous? – silence of the English and French about the book could be another way to express that “something happened.” For once, humanist scholars take their time, what should be always the case when translation is at stake.

I would have preferred that the German subtitle evoke the figure of a “digital Meister,” a *digitaler Meister in den Literatur- und Kulturwissenschaften*. Indeed, Wenders and Goethe are both narrating the initiation of a young man, Wilhelm Meister. The Wilhelm Meister of Wenders is less successful than the one of Goethe. Travelling through Germany – still divided in 1975 – Wenders’ Meister expresses his melancholia at the end of the movie, saying: “Avevo l’impressione di aver mancato qualcosa, e di continuare a mancare qualcosa, a ogni nuovo movimento”; “I had the impression to have missed something, and to continue missing something at each new movement”.<sup>23</sup> It is not obvious to qualify such travel as a *Wende*, a peaceful “turn-around” process. Almost fifty years after the movie of Wenders, our present “digital Meister” faces a world tempted again by division, even if transformed by digitization.

---

<sup>17</sup> Moretti, *Falso*, 24: “Il testo era al centro dello studio della letteratura: in questo caso, e un puntino. È stato ridotto a un puntino”; we arrived to “una storia della letteratura senza testi. Senza testi – nel senso che ce ne sono troppi, naturalmente, e quindi non li si può più studiare *nella loro individualità*.”

<sup>18</sup> Moretti, *Falso*, 36: “L’interpretazione si muove *tra la forma et il mondo*, andando in cerca del significato storico delle opere; la quantificazione si muove *tra forma e forma*, tentando di tracciare le coordinate di un futuro atlante della letteratura. Per la prima, la forma è *une forza, un agire*, un modo di ‘plasmare’ l’esistente che va accolto con sospetto, contrastato, e infine smascherato. Per l’altra, la forma è un *prodotto finito*: da misurare a mente fredda, e collocare all’interno di un vasto sistema di rapporti.”

<sup>19</sup> Moretti, *Falso*, 99.

<sup>20</sup> Moretti, *Falso*, 127.

<sup>21</sup> See <https://simple.wikipedia.org/wiki/Wende>.

<sup>22</sup> See for example Andrew Piper and Mark Algee-Hewitt, “The Werther Effect I: Goethe, Objecthood, and the Handling of Knowledge” (2014); Da (2019, 633-634); Piper (2020, 7; in answer to Da); Primorac et al. (2023, 1).

<sup>23</sup> Moretti, *Falso*, 9-10.

#### 4. CONCLUSION: AT THE HORIZON

Challenging the quite successful initiation of Wilhelm Meister of Goethe remains a task of European literature. A detail went unmentioned by Moretti: in Goethe's work, Wilhelm reads Shakespeare and plays Hamlet. At the end of this conference paper, we reach the missing point: the English language and culture. Hamlet has inspired Wilhelm by Goethe, who inspired Wilhelm by Wenders, who inspired Moretti in *Falso movimento*. Logically, it means that "tôt ou tard", sooner or later, English language scholarship should consider Moretti's *Falso movimento*. Translators and publishers will surely be encouraged by the positive insights also present in the book, culminating with this statement: "The day the abundance of data will open the road to ambitious concepts – instead of discouraging them – this day, the quantitative history of the culture would have truly started, and the confrontation with the 'other' humanist disciplines will finally become possible."<sup>24</sup> Will the more positive Wilhelm of Goethe encourage the melancholic Wilhelm of Wenders to pursue the digital humanities distant reading adventure coined by Moretti? A modest but clear answer can be given to this question: whatever happens, this task will be probably performed by the next generation, those who still did not make too many *wrong moves*.

#### REFERENCES

- [1] Bardiot, Clarisse et al. *La fabrique numérique des corpus en sciences humaines et sociales*. Presses Universitaires Septentrion. Paris, 2022.
- [2] 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>
- [3] Da, Nan Z. "The Computational Case against Computational Literary Studies", *Critical Inquiry* 45 (2019/3), 601-638, <https://www.journals.uchicago.edu/doi/10.1086/702594>
- [4] Gallegati, Sara. "Moretti, Franco 2022. *Falso movimento*. La svolta quantitativa nello studio della letteratura. Milano: notttempo", *Umanistica Digitale* 13 (2022): 153-156.
- [5] Geulen, Eva. "DISTANT READING UP CLOSE: Moretti zieht Bilanz", *ZfL Blog* (25.11.2022): 1-8, <https://doi.org/10.13151/zfl-blog/20221125-01>
- [6] Marra, Demetrio. "Sul Falso movimento di Franco Moretti (e cioè sull'intera collana, se vogliamo)". *Lingua Italiana* (20.07.2022). [https://www.treccani.it/magazine/lingua\\_italiana/articoli/percorsi/percorsi\\_394.html](https://www.treccani.it/magazine/lingua_italiana/articoli/percorsi/percorsi_394.html)
- [7] de Maeyer, Juliette. "Travailler les corpus de presse numérique, un travail d'artisan?", in Clarisse Bardiot et al., *La fabrique numérique des corpus en sciences humaines et sociales*, Presses Universitaires Septentrion. Paris, 2022: § 1-30.
- [8] Moretti, Franco. *La letteratura vista da lontano*. Giulio Einaudi. Torino, 2005.
- [9] Moretti, Franco. *Graphs, Maps, Trees. Abstract Models for a Literary History*. Verso. London / New York, 2005.
- [10] Moretti, Franco. *Graphes, cartes et arbres. Modèles abstraits pour une autre histoire de la littérature*. Les prairies ordinaires. Paris, 2008.
- [11] Moretti, Franco. *Distant Reading*. Verso. New York, 2013.
- [12] Moretti, Franco. *Falso Movimento*. La svolta quantitativa nello studio della letteratura. notttempo. Milano, 2022.
- [13] Moretti, Franco. "Falso Movimento. La svolta quantitativa nello studio della letteratura". *Le parole e le cose* 2 (21.01.2022). <https://www.leparoleelecose.it/?p=43269>
- [14] Moretti, Franco. *Falsche Bewegung*. Die digitale Wende in den Literatur- und Kulturwissenschaften. Konstanz University Press. Konstanz, 2022.
- [15] Piper, Andrew and Algee-Hewitt, Mark. "The Werther Effect I: Goethe, Objecthood, and the Handling of Knowledge", in Matt Erlin and Lynn Tatlock (eds), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Camden House. Rochester, 2014: 155-184.
- [16] Piper, Andrew. "Do We Know What We Are Doing?", *Journal of Cultural Analytics* 5 (2020/1), 1-13, <https://culturalanalytics.org/article/11826>
- [17] Primorac, Antonija et al. "Distant Reading Two Decades on: Reflections on the Digital Turn in the Study of Literature", *Digital Studies / Le Champ Numérique* (2023), accepted version, <https://ora.ox.ac.uk/objects/uuid:acbba668-71d1-4f8e-9cfc-fda801214682>.
- [18] Soscia, Danilo. "La svolta quantitativa spiegata a mio figlio. A proposito di 'Falso movimento' di Franco Moretti", *minima&moralia* (14.02.2022), <https://www.minimaetmoralia.it/wp/libri/la-svolta-quantitativa-spiegata-a-mio-figlio-a-proposito-di-falso-movimento-di-franco-moretti/>
- [19] Terrisse, Bénédicte and Wögerbauer, Werner. "La carte et la frise: des métaphores aux diagrammes", *Germanica* 71 (2022), 7-18, <https://doi.org/10.4000/germanica.18470>

---

<sup>24</sup> Moretti, *Falso*, 127: "Il giorno che l'abbondanza di dati aprirà la strada a concetti ambiziosi, invece di scoraggiarli – quel giorno, la storia quantitativa della cultura avrà davvero inizio, e il confronto con le 'altre' discipline umanistiche diventerà finalmente possibile."

# Verso la creazione della versione italiana di SPARSAR

Rodolfo Delmonte<sup>1</sup>, Francesco Stiffoni<sup>2</sup>

<sup>1</sup> Università Ca' Foscari Venezia, Italia – delmont@unive.it

<sup>2</sup> Università Ca' Foscari Venezia, Italia – fstiffo@sgajo.com

## ABSTRACT

In questo lavoro presentiamo il lavoro in corso per la creazione della versione italiana di SPARSAR sistema creato per l'analisi e la visualizzazione del contenuto poetico – quindi, prosodico, retorico e semantico – di poesie inglesi a partire dalla poesia elisabettiana. Il lavoro di conversione ha attualmente raggiunto e completato un primo traguardo che permette al sistema di leggere attraverso la sintesi vocale disponibile su computer la poesia analizzata con sufficiente espressività e sono in fase di applicazione le componenti che sono responsabili per una manipolazione accurata del contorno intonativo. Descriveremo quindi quanto già funzionante e il modulo che si preoccupa di costruire il contorno intonativo sulla base della struttura creata dalla interazione tra fonologia, sintassi, semantica e pragmatica.

## PAROLE CHIAVE

Analisi computazionale del testo; analisi poetica; analisi prosodica; analisi fonologica; TTS; voci e modelli neurali; liste di frequenza; modello ToBI

## 1. INTRODUZIONE

In questo lavoro presentiamo le attività in corso per la creazione della versione italiana di SPARSAR - System for Poetry Automatic Rhythm and Style AnalyzeR (la cui prima presentazione è stata nel 2013[8])– sistema creato per l'analisi e la visualizzazione del contenuto poetico – quindi, prosodico, retorico e semantico – di poesie inglesi a partire dalla poesia elisabettiana[4]. Il lavoro di conversione ha attualmente raggiunto e completato un primo traguardo che permette al sistema di leggere attraverso la sintesi vocale disponibile su computer la poesia analizzata con sufficiente espressività e sono in fase di applicazione le componenti che sono responsabili per una manipolazione accurata del contorno intonativo[5]. La versione definitiva del sistema avrà come risultato un algoritmo dedicato a valutare la combinazione degli espedienti poetici e figure retoriche, nonché delle valenze semantiche e pragmatiche che possono convergere oppure no per creare un unico flusso di combinazioni senso-suono[7].

Descriveremo ora quanto già funzionante e il modulo che si preoccupa di costruire il contorno intonativo sulla base della struttura creata dalla interazione tra fonologia, sintassi, semantica e pragmatica.

In particolare il sistema produce le seguenti analisi del testo di una poesia:

- analisi morfologica e tagging grammaticale;
- analisi sintattica a costituenti e a dipendenze;
- analisi semantica a livello di parola e di enunciato per l'indicazione del “sentiment”;
- ricerca delle parole rare sulla base delle liste di frequenze disponibili, quella chiamata WaCKy<sup>1</sup> – 450mila parole ricavata dal corpus PAISA<sup>1</sup> di 250M di tokens - confrontata con altre liste – tra cui quella di SUBTLEX<sup>2</sup> fatta di 500mila parole da 50M di testi, quella di ColFIS<sup>3</sup> – 173mila parole da un corpus di 3.5M di testi, quella di Buchmeier<sup>4</sup> – 75mila parole da un corpus di 5.5M di testi, per mostrare la variabilità legata alla tipologia dei testi che compongono il corpus di riferimento[2,3];
- analisi fonologica che include la trascrizione fonetica, la separazione in sillabe, e l'assegnazione dell'accento di parola
- traduzione del risultato dell'analisi precedente in due formati: un primo formato in ARPAbet e in IPA con l'aggiunta però delle informazioni prosodiche, separazione in sillabe e accento di parola; un secondo formato per la lettura da parte del sintetizzatore disponibile su computer Apple. Il primo formato serve per introdurre nel secondo formato parole che il sintetizzatore non sarebbe in grado di leggere correttamente – ne parlo in seguito.

---

1 <https://wacky.sslmit.unibo.it/>

2 <http://crr.ugent.be/programs-data/subtitle-frequencies>

3 <https://www.istc.cnr.it/en/grouppage/colfis>

4 [http://en.wiktionary.org/wiki/User:Matthias\\_Buchmeier](http://en.wiktionary.org/wiki/User:Matthias_Buchmeier)

- traduzione nel formato SSML, il markup language utilizzato dalla maggior parte dei sintetizzatori vocali disponibili online, tra cui quello della Microsoft nell'ambiente chiamato Azure<sup>5</sup>, o quello di Polly<sup>6</sup> di Amazon.
- costruzione del contorno intonativo e della struttura prosodica complessiva.

Abbiamo testato il sistema utilizzando poesie scritte da poeti a cavallo tra l'800 e il '900. La lingua italiana, soprattutto quella usata nella poesia dell'800 e dell'inizio del '900, da poeti come Foscolo, Carducci, Pascoli e altri utilizza figure retoriche e poetiche tipiche, in parte ereditate dai grandi poeti del '200 e del '300, Dante, Petrarca e tramandate a quelli del '400 e '500 come Ariosto. Le strutture di frase sono la sede per realizzare effetti essenziali a livello semantico che si riflettono sulla struttura prosodica e su quella fonetica e fonologica, e sono frequentemente non-canoniche[6]. Ad esempio, nella incipit della poesia “Il pioppo” di Clemente Rebora, – che utilizziamo per mostrare come opera SPARSAR il cui testo si trova in Appendice – il soggetto viene dislocato alla fine della frase nel verso successivo al primo che comincia con il verbo “Vibra”, parola sinestetica fortemente evocativa dal punto di vista sensoriale. Il terzo verso ripropone lo stesso schema ma con maggior forza, al pioppo viene attribuita un’anima – “spasima l'anima”, una sequenza di due parole sdrucchiole con l'accento posto sulla stessa vocale “a”, costruita di nuovo come soggetto in posizione post-verbale. Il primo verso della seconda strofa presenta di nuovo una dislocazione, in questo caso l'anteposizione degli aggiunti, tre sintagmi preposizionali a cui fa seguito il verbo reggente che di nuovo ripropone un pioppo umanizzato che “si esprime”. Questo verso è seguito da una frase participiale modificatrice del terzo aggiunto preposto la cui testa “fronde” si accorda con il participio “tese”. E di nuovo abbiamo una struttura non-canonica, in cui il participio è preceduto dall'aggiunto “al ciel”, suo dipendente, e questo a sua volta è preceduto dal quantificatore “tutte” che nella versione “floating” dà un afflato rituale ed è oggetto di varie ripetizioni. Il verso successivo inizia con un predicato aggettivale preposto e una inversione argomentale completa con il soggetto di nuovo posposto. In questa seconda strofa la parola “tronco” viene ripetuta tre volte, legando assieme i relativi versi da cui sappiamo che è “del mistero” e che “s'inabissa” con le sue radici nel “vero”. Per tutte queste strutture sintattiche e metaforiche il contorno intonativo dovrebbe restituire variazioni significative ed espressive adeguate. Il contorno intonativo costituisce il cuore del sistema ed è responsabile per la maggior parte del contenuto “espressivo” prodotto. La base teorica che utilizziamo per la costruzione del contorno intonativo è ToBI – Tones and Break Indices[1,9]. Il modulo che è in fase di completamento ha a disposizione l'informazione fonologica e prosodica a livello di parola, quella sintattica funzionale per l'individuazione delle teste e degli eventuali dipendenti, quella semantica per le variazioni del contorno intonativo e agisce e agirà nel modo seguente:

- creare la griglia di base per la struttura intonativa partendo dalle parole che possono ricevere i Pitch Accents;
- sulla base dell'informazione sintattica e funzionale inizio della creazione dei Gruppi Intonativi distinguendo le teste dai dipendenti e introdurre ulteriori marche per i Phrase Accents;
- sempre a livello sintattico si interviene sull'inserimento di break nelle posizioni motivate dalla sintassi ed unisce le parole in presenza di enjambement a livello di verso e di strofa, completando i Gruppi Intonativi;
- a livello semantico si decidono le variazioni del contorno intonativo a livello di enunciato per produrre l'Accento di Frase, differenziando strutture argomentali canoniche da quelle non-canoniche aggiungendo enfasi nel secondo caso;
- sempre a livello semantico – in parte con il “sentiment” e in parte con il calcolo dei tratti aspettuali dell'enunciato - si differenziano le componenti del contorno intonativo che sono in “foreground” da quelle che invece sono in “background”
- quello pragmatico basato sull'analisi del lessico di frequenza e della struttura retorica e poetica si aggiungono effetti locali – modulando a livello sillabico parole rilevanti – questa ultima parte non è ancora stata elaborata anche se le informazioni sono tutte disponibili.

Nella sezione successiva illustriamo lo stato dell'arte nel campo del TTS spiegando la peculiarità dell'approccio corrente basato sulla voci neurali e sul problema che hanno i modelli con le parole rare che nella poesia italiana costituiscono una percentuale importante. Il testo finale è accompagnato da alcuni file .wav che servono a documentare la variazione espressiva che si ottiene partendo dalla lettura di base a quella finale che utilizza i parametri prodotti dal sistema, di cui forniremo il corrispettivo file in formato ssml.

<sup>5</sup> <https://azure.microsoft.com/it-it>

<sup>6</sup> <https://aws.amazon.com/it/polly/>



## 2. LO STATO DELL'ARTE NEL CAMPO DELL'ANALISI AUTOMATICA

Il lavoro computazionale sulla poesia indirizza un gran numero di discipline che sono comunque fortemente correlate. Queste includono l'annotazione linguistica manuale e automatica, l'analisi linguistica o psicolinguistica e cognitiva, o la traduzione della poesia, come pure la sua creazione o generazione automatica, che commentiamo qui di seguito. Un primo sotto campo specifico riguarda la traduzione grafema-fonema per tutte le parole fuori vocabolario, come discusso in [10,11] per tutte le applicazioni che usano un vocabolario che riporta la pronuncia, operazione fondamentale per derivare la posizione dell'accento e della rima; la traduzione è inoltre fondamentale per incorporare restrizioni sulla struttura metrica e lo schema di rima utilizzabile da sistema di Machine Translation e classificatori a Deep Neural Networks. A questo scopo c'è stato diverso lavoro a livello computazionale per caratterizzare lo schema delle rime [12] e le proprietà globali delle strutture ritmiche in inglese [13]. A questo scopo Green et al. [14] usano un trasduttore a stati finiti per inferire l'assegnazione dell'accento a livello di sillaba in ambito poetico che utilizza restrizioni metriche. Utilizzano metodi statistici per analizzare, generare e tradurre poesia. Dapprima applicano procedure di apprendimento non-supervisionato per rivelare gli schemi accentuali in un corpus di poesia in forma ortografica. Successivamente, utilizzano questi schemi di accento di parola, in aggiunta a modelli di rime e di discorso, per generare poesia inglese d'amore. Alla fine traducono poesia italiana in inglese, scegliendo realizzazioni finali che si adattano agli schemi ritmici desiderati. Però si concentrano su un solo tipo di metro poetico, il pentametro giambico.

Esiste un numero importante di lavori sulla generazione di poesia a cominciare da paper di R.Tsur e P.Gérvás [15,16,17], che usa Case Based Reasoning (Ragionamento Basato su Esempi) che sembra indurre la migliore struttura del verso. Altri tentativi interessanti sono quelli di Toivanen et al.[18] che usa un approccio basato su corpora per generare poesia in finlandese. La loro idea è di contribuire la conoscenza necessaria sia nel contenuto che nella forma attraverso i loro corpora separati, uno che fornisce il contenuto semantico, e l'altro quello grammaticale e quello poetico. A questo scopo uniscono analisi morfologica e sintesi assieme a metodi di text-mining.

La generazione di poesia in basco è l'argomento di lavori prodotti da Agirrezabal et al. [19], i quali utilizzano il POS-tagging (o etichettatura per categorie grammaticali) per indurre l'ordinamento lineare delle parole e WordNet per selezionare la scelta semantica migliore nel contesto dato.

Manurung et al., [20,21] hanno esplorato il problema della generazione di poesie sotto restrizioni, utilizzando tecniche di Machine Learning. Con il loro lavoro, gli autori intendevano riempire il vuoto esistente nel paradigma generativo, e "di far luce su quello che spesso sembra la più enigmatica e misteriosa forma di espressione artistica". La conclusione che raggiungono è che "malgrado la nostra implementazione sia in uno stadio molto iniziale, il risultato ottenuto mostra chiaramente che modello di ricerca stocastico a "hillclimbing" riesce a produrre testo che soddisfa le restrizioni imposte". Reddy & Knight [22] producono un algoritmo di apprendimento automatico non-supervisionato per trovare schemi di rime che si suppone indipendente dalla lingua. Il loro approccio lavora sull'intuizione che ogni raccolta di poesia in rima contiene inevitabilmente ripetizioni di coppie in rima. Questo fatto nella lingua inglese è in parte dovuto alla "sparsity" (scarsità) dei modelli di rima, inoltre molte parole non hanno rima e molte altre ce l'hanno ma solo in misura minima, obbligando i poeti a riutilizzare le stesse coppie. Sulla base di questo approccio, gli autori creano un algoritmo non-supervisionato per inferire gli schemi di rima basandosi su un modello di generazione a livello di strofa. La definizione di rima che gli autori usano è quello della rima baciata o perfetta, altri tipi di rima non vengono considerati. I risultati ottenuti dall'algoritmo vengono controllati dal dizionario fonologico CELEX [23].

Esiste un numero limitato di sistemi basati su regole e disponibili per il download che elenco partendo dal più vecchio a quello più recente:

- lo Scandroid by C.Hartman (2004/5), scaricabile a <http://oak.conncoll.edu/cohar/Programs.htm>, e presentato in [24]
- lo Stanford Literary Lab di Algee-Hewitt, M., Heuser, R., Kraxenberger, M., Porter, J., Sensenbaugh, J., e Tackett, J. (2014), downloadabile a <https://github.com/quadrismegistus/litlab-poetry>, e presentato in [25,26]
- il progetto Canadian Representative Poetry Online, della University of Toronto realizzato da M.R. Plamondon e documentato in [27], scaricabile nella Library website <http://rpo.library.utoronto.ca/>
- uno sforzo collaborativo prodotto da università americane e tedesche chiamato MYOPIA, presentato in [28] e disponibile in due websites dalla autrice principale Helen Armstrong, <https://lecture2go.uni-hamburg.de/konferenzen/-/k/13930>, <http://www.helenarmstrong.us/design/myopia/>
- ZeuScansion uno strumento per la scansione della poesia English realizzato da M. Agirrezabal et al. e presentato in [29], e disponibile a <https://github.com/manexagirrezabal/zeuscansion>

- RhymeDesign uno strumento disegnato per l'analisi di dispositivi metrici e ritmici, da N.McCurdy et al. [30]; lo strumento era precedentemente chiamato Poemage, documentato a <http://www.sci.utah.edu/~nmccurdy/Poemage/> e ora presentato come progetto a [http://ninamccurdy.com/?page\\_id=398](http://ninamccurdy.com/?page_id=398)
- Un certo numero di lavori più o meno recenti hanno indirizzato il problema relativo alla identificazione della rima, ad esempio da Manish Chaturvedi et al. [31], e da Karteek Addanki e Dekai Wu [32], e precedentemente da Hussein Hirjee and Daniel Brown [33] e da Susan Bartlett et al. [34].
- Per finire una lista selezionata di autori hanno indirizzato specificamente il problema della visualizzazione di dati linguistici and letterari in lavori più o meno recenti, in particolare sulla visualizzazione della poesia [35-40], l'analisi letteraria e le concordanze [41-43].

Nessuno dei sistemi indicati lavora per la lingua italiana o lingue simili. Recentemente un progetto europeo cerca di produrre un sistema simile a SPARSAR per la lingua spagnola: si trova documentato a <https://postdata.linhd.uned.es/> e documentato a <https://github.com/linhd-postdata>, dove l'acronimo sta per Poetry Standardization and Linked Open Data. È inteso come piattaforma digitale per attività sulla poesia, indirizzata sia a scolari, accademici e utenti naive. Il progetto è molto ambizioso e attualmente è in fase iniziale<sup>7</sup>.

### 3. IL TTS: DAI DIFONI ALLE RETI NEURALI

La maniera tradizionale di lettura di un testo da parte di un sintetizzatore vocale - in breve TTS - è rappresentata nella figura 1. dove abbiamo unito la sequenza di moduli utilizzata dalla sintesi per concatenazione di difoni a quella attuale realizzata grazie alle DNN per cui si passa da una sequenza di caratteri ortografici all'output in forma di onda vocale. Come si può notare, il modulo linguistico ha il compito di creare i parametri acustici che serviranno al TTS per produrre il parlato. Questo modello è stato utilizzato dall'inizio degli anni '60 fino a una ventina di anni fa quando hanno iniziato ad entrare in funzione i cosiddetti voice assistants e le virtual talking heads o avatars. A questo punto c'è stata una spinta per realizzare prodotti che fossero adatti e adattabili facilmente ad ogni esigenza e soprattutto ad ogni lingua. Questo fatto ha provocato la fine della sintesi concatenativa per difoni ma anche quella promettente ma più difficile che associava parametri con una sintesi articolatoria. Questo fatto viene considerato negativamente da R.Sproat e J. van Santen che esprimono il loro giudizio dicendo che il TTS non è ancora una tecnologia matura. Quello che è avvenuto successivamente si basa esclusivamente su procedure statistiche, dapprima guidate dai HMMs (Hidden Markov Models) e successivamente dalle Deep NNs (Neural Networks).

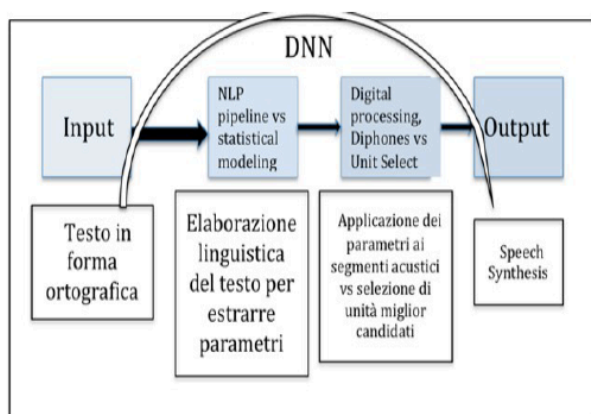


Figura 1, Schema processi linguistici di un TTS

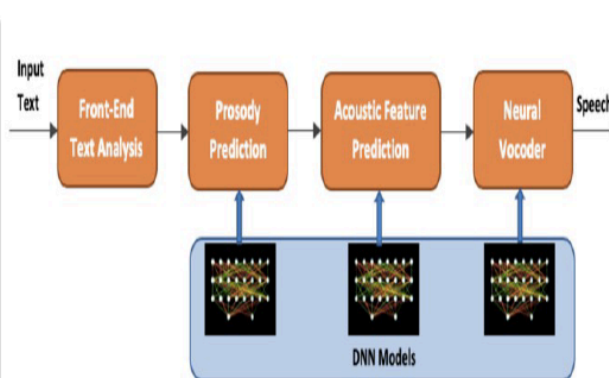


Figura 2, Il modello del TTS a DNNs della IBM

L'approccio statistico richiede nella prima fase che i segmenti residenti su database siano opportunamente e accuratamente annotati in precedenza (Delić et al. 2017). Ma utilizzando segmenti restano insoluti tutti i fenomeni di coarticolazione per coprire i quali il numero di segmenti da memorizzare sarebbe troppo elevato e la loro funzione perderebbe significato a causa del fenomeno della "sparseness" (scarsità) dei dati.

I sistemi TTS correnti invece vogliono saltare dal segmento all'unità parola o sequenza di parole (phrase) per poter inglobare i fenomeni di coarticolazione. Si parte in questo caso dalla forma d'onda vocale del parlato con tutti i suoi parametri in

<sup>7</sup> Ci sono purtroppo problemi nel database projectgutenberg impostato su Github che dovrebbe contenere copia di tutta la poesia universalmente disponibile in forma digitale. Il file di italian\_poetry è corrotto e riporta la *Divina Commedia* tre volte di seguito.

forma di spettrogramma e il modello creato viene poi usato per predire e generare con le DNNs altri spettrogrammi che vengono uniti per produrre il parlato (Shen 2017; Wang 2018). Mostriamo in Figura 2. il modello a DNN adottato dall'IBM che comunque replica modelli simili per altri TTS attuali.

In questi approcci stocastici, la prosodia deve essere modellata separatamente e aggiunta all'inizio del passo di generazione dell'onda sonora. La prosodia richiede un nuovo modello predittivo che può essere creato di nuovo sulla base di processi probabilistici. Comunque questo componente è quello che risulta essere il più difficile a causa della variabilità presente negli enunciati del parlato naturale. Il sistema TTS di Google, chiamato TACOTRON 2, usa una procedura simile, ma la conclusione degli autori è la seguente “We’d also like to develop techniques to select appropriate prosody or speaking style automatically from context, using, for example, the integration of natural language understanding with TTS.”

<https://www.googblogs.com/tag/tts/> e

[https://google.github.io/tacotron/publications/end\\_to\\_end\\_prosody\\_transfer/](https://google.github.io/tacotron/publications/end_to_end_prosody_transfer/)

Abbiamo testato diversi sistemi ma in genere la organizzazione del contorno intonativo è legato alla presenza di punteggiatura. Azure della Microsoft ha voci sintetiche molto naturali ma non capaci di pronunciare correttamente le parole troncate o apocope, a differenza del TTS di Nuance. Questo fatto è ovviamente comprensibile visto che non si trovano nei dizionari dell'italiano attuale e visto che nei database di parlato è impossibile trovare letture di poesie dell'800. Il problema è che mancano anche tutte le spezzature o scavalcamenti o semplicemente "enjambements" che richiedono da parte del TTS di capire che la frase o il sintagma non è completo a fine verso e che bisogna congiungerlo con il verso successivo, anche se si trova in una strofa separata. Invece quello che si sente è il posizionamento di una pausa e di una caduta o "declination" del contorno intonativo invece di una "continuation". Per verificarlo basta collegarsi al sito di Nuance dove è possibile utilizzare una delle voci italiana disponibili e inserire il testo di una poesia.

<https://www.nuance.com/omni-channel-customer-engagement/voice-and-ivr/text-to-speech.html#!>

#### **4. I MODULI PRODOTTI E L'ANALISI DELLA POESIA “IL PIOPPO”**

Indichiamo ora alcune delle principali difficoltà incontrate da SPARSAR<sup>8</sup> e un esempio di analisi sulla base della poesia “Il pioppo” di Clemente Rebora. Un primo problema da risolvere sono le parole apocope, che cioè hanno subito l'elisione della vocale atona della sillaba finale come ad esempio in “pensier” da “pensieri” o “furon” da “furo”, “uom” per “uomo”, ma anche parole arcaiche e in disuso come “godea” per “godeva”, o “deggio” per “dovere”. Le parole troncate devono essere pronunciate rispettivamente con l'accento sulla seconda e ultima sillaba dopo il troncamento, “sier” la prima e con l'accento sulla prima sillaba “fu” la seconda. In pratica le parole troncate mantengono la posizione dell'accento della forma intera. Lo stesso vale per la caduta della consonante fricativa labiodentale sonora in posizione intervocalica. Ovviamente per saperlo è necessario ricomporre la parola prima di analizzarla. Questo passaggio è essenziale per tutta l'analisi linguistica che segue, quella sintattica e semantica. Parole come “cor” che riprende la forma latina di “cuore” vengono trasferite in forma intera modificata nel “parser” che farà l'analisi sintattico-semantica. Parole rare come “rinverdiva”, “illacrimata”, “ritrosi”, “oblioso” ecc. di cui il sistema non possiede il lemma né la forma, e che quindi costituiscono parole fuori vocabolario (indicate come “OutOfVocabularyWords” OOVW in letteratura di ASR – Automatic Speech Recognition), vengono trattate dall'analizzatore morfologico e sulla base del loro suffisso assegnate a una categoria lessicale Verbo Flesso, Verbo Partecipio Passato, Aggettivo le ultime due. L'analisi morfologica e quindi il tagging è indispensabile nella fase di trascrizione fonemica per l'assegnazione dell'accento di parola e non solo, considerando che l'italiano possiede una quantità indefinita di parole omografe non omofone, e che soprattutto ha fenomeni di cliticizzazione che richiedono di nuovo di posizionare l'accento nella posizione della parola priva di clitici, come in “ricordatemelo”, “facciamocelo” o “fammelo”. Parole come queste sono tipicamente assenti dai vocabolari e devono quindi essere ricostruite per regole, ma lo possono fare sapendo che sono state taggate come verbi che contengono clitici e che prima di essere trascritte foneticamente devono essere separate dal clitico che poi viene ricongiunto ad accento e sillabazione avvenuta.

##### **4.1. Analisi della poesia “il pioppo”**

Mostriamo ora l'analisi della poesia “Il pioppo” di Clemente Rebora che il sistema legge direttamente su computer Apple con una propria codifica interna, ma che contemporaneamente fornisce una traduzione in linguaggio SSML per qualsiasi altro sistema di TTS. I dettagli dell'analisi verranno riportati nella versione finale del paper. Consideriamo per ora solo il fatto che alla costruzione della struttura intonativa arriverà tutta l'informazione costruita precedentemente, a cominciare dall'elenco delle parole rare, delle parole che comportano un “sentiment” negativo, la struttura a costituenti e quella

---

<sup>8</sup> Il sistema è realizzato in linguaggio Prolog e viene distribuito gratuitamente per ricerca o studio dal suo sito web, <https://sparsar.wordpress.com/sparsar/> o direttamente contattando l'autore per email.

predicato-argomentale, il risultato dell'analisi aspettuale e semantica a livello proposizionale, nonché le informazioni sulla struttura poetica, le parole di fine verso e la separazione in strofe, ecc.

La struttura intonativa viene creata frase per frase considerando uguale a 100 la sua lunghezza in caratteri fonetici e segnando in percentuale la posizione di ciascun pitch accent prima sulla base della posizione della parola nella frase e poi sulla base della posizione nella sillaba tonica.

Nell'analisi fonologica utilizzerò la rappresentazione fonetica interna che non richiede l'utilizzo di simboli particolari nel word processor OpenOffice di cui mi servo. Questa trascrizione utilizza i seguenti caratteri:

- k – per la oclusiva sorda velare |co, ca, ch|
- x – per la fricativa dentale sonora |s|
- X – per la fricativa palatale sorda |sc|
- % - per la oclusiva palatale sonora |ge, gi|
- G – per la oclusiva velare sorda |go, ga, gh|
- < - per la laterale palatale |gl|
  - - per la nasale palatale |gn|
- \$ - per la affricata dentale sonora |z|
- & - per la vocale E aperta
- @ - per la vocale O aperta
- j – per l'approssimante palatale I |iu, io, ia, ie|
- w – per l'approssimante velare U |uo, ua, ui, ue|

#### 4.2. La struttura intonativa – prima fase

Questa è la struttura intermedia che contiene la traduzione fonemica e prosodica e la indicazione dei “pitch accents” del contorno intonativo in percentuale. La struttura finale intonativa conserverà le percentuali delle teste ed eliminerà quelle delle parole dipendenti. Ogni percentuale verrà poi associata a un numero che esprime l'altezza del contorno intonativo in Hertz come previsto dalla piattaforma Azure di Microsoft, o in Polly di AWS.

[il 29%-pioppo-pj@1\_ppo di 71%-Clemente-kle\_m&n1\_te 92%-Rèbora-r&1\_bo\_ra],

[9%-vibra-vi1\_bra nel 26%-vento-v&1\_nto con 43%-tutte-tu1\_tte le sue 67%-foglie-fo1\_<e il 85%-pioppo-pj@1\_ppo 98%-severo-se\_v&1\_ro],

[9%-spàsima-spa1\_xi\_ma 25%-l-anima-la1\_ni\_ma in 38%-tutte-tu1\_tte\_ le sue 58%-dòglie-d@1\_<e 68%-nell-ansiane\_lla1\_nsj del 83%-pensiero-pe\_nsj&1\_ro:],

[dal 3%-tronco-tron1\_ko in 16%-rami-ra1\_mi per 27%-fronde-fr@n1\_de si 40%-esprime-e\_spril\_me 51%-tutte-tu1\_tte al 63%-ciel-c&l1 68%-tese-te1\_se con 79%-raccolte-ra\_kkoll\_te 92%-cime-ci1\_me],

[2%-fermo-fer1\_mo]

[7%-rimane-ri\_ma1\_ne il 20%-tronco-tron1\_ko del 36%-mistero-mi\_st&1\_ro, e il 63%-tronco-tron1\_ko 80%-s-inabissasi\_na\_bi1\_ssa 86%-ov-è-o1\_v& 93%-più-pju1 98%-vero-ve1\_ro]

Il completamento della struttura intonativa deve però tenere in considerazione la suddivisione in strofe e per ciascuna strofa la suddivisione in versi che viene fatta da questa struttura poetica in cui viene indicata la strofa – due segnali di strofa servono per separate il titolo e l'autore dalla poesia - e la parola fine verso:

[ 1-([pioppo-'pj@1\\_ppo'](#)), 2-([Rèbora-'r&1\\_bo\\_ra'](#)),

3-strofa,

4-([foglie-'fo1\\_<e'](#)), 6-([severo-'se\\_v&1\\_ro'](#)),

6-strofa,

7-([dòglie-'d@1\\_<e'](#)), 9-([pensiero-'pe\\_nsj&1\\_ro'](#)),

9-strofa,

12-([esprime-'e\\_spril\\_me\\_'](#)),14-([cime-'ci1\\_me'](#)),

14-strofa,

15-([mistero-'mi\\_st&1\\_ro'](#)),17-([vero-'ve1\\_ro\\_'](#))]'

Dalla semantica si ottiene l'elenco della parole con un “sentiment” marcato che sono le seguenti:

[spasima,doglie,ansia,inabissa].

Le strutture non-canoniche vengono individuate sulla base della struttura a costituenti con marche funzionali in cui la marca di soggetto a livello di frase è stata segnata come OGGetto mentre nell'analisi a dipendenze e semantica la stessa testa appare con la marca SOGGetto. Queste strutture vengono marcate con enfasi, come anche le parole rare che in questa poesia si limitano a “Vibra”, “inabissa”.

L'andamento del contorno intonativo complessivo è deciso dalla struttura informativa e dalla rilevanza che viene calcolata nell'analisi semantica mediante la scelta tra due etichette: “Foreground” e “Background”. Nella poesia queste sono le marche associate ai predicati verbali di cui si indica anche il lemma e la posizione nel testo:

[ 6-'Vibra'-vibrare-foreground,  
18-spasima-spasimare-background,  
38-esprime-esprimere-foreground,  
42-tese-tendere-background,  
44-raccolte-raccogliere-background,  
48-rimane-rimanere-foreground,  
58-inabissa-inabissare-foreground]

### 4.3. La struttura intonativa – seconda fase

Nella seconda fase viene creato il contorno intonativo a livello di enunciato, accompagnando le posizioni dei pitch accents individuati precedentemente con un valore in Hz come viene mostrato qui di seguito. La marca utilizzata seguendo il documento pubblicato sul sito WWW3 (<https://www.w3.org/TR/speech-synthesis11/#ref-ipahndbk>) è <prosody contour> il quale prevede che ogni segmento o enunciato sia caratterizzato da una indicazione della velocità di eloquio con la marca <rate> e uno dei possibili tipi previsti, oppure una indicazione in msec. Inoltre la struttura prosodica può contenere delle eventuali pause aggiuntive rispetto a quelle previste dalla punteggiatura, indicate queste con la marca <break> e una durata in msec., e una indicazione del livello di volume o intensità o energia in dB con la marca <volume>. Un ulteriore elemento può essere aggiunto – e come detto più sopra noi lo prevediamo per parole rare in strutture non-canoniche – ed è l'enfasi, con la marca <emphasis> e alcune etichette di livello. Riportiamo quindi come esempio quanto prodotto da SPARSAR per la poesia “Il Pioppo”.

1. Il pioppo<prosody contour="(13%, -54%" rate="slow">di Clemente Rèbora</prosody>.
- 2.<prosody contour="(9%, +57%) (12% +42%" rate="x-slow"><emphasis level="strong">Vibra,</emphasis></prosody>
- 3.<prosody contour="(26%, +17%) (41%, -21%) (65%, +50%) (95% -40%" rate="+10%">nel vento con tutte le sue foglie</prosody>
4. <prosody contour="(45%, -32%) (93%, -44%" rate="slow">il pioppo severo;</prosody>
- 5.<prosody contour="(26%, +17%) (41%, -21%) (65%, +50%) rate="x-slow">spasima l'anima</prosody>
- 6.<prosody contour="(26%, +17%) (41%, -21%) (65%, +50%" rate="x-slow">in tutte le sue doglie</prosody><break strength="medium"/>
- 7.<prosody contour="(20%, -68%) (55%, -50%) (85%, -67%" rate="slow">nell'ansia del pensiero:</prosody>
- 8.<prosody contour="(16%, +54%) (41%, +73%) (63%, +14%) (91%, +41%) (98%, +75%" rate="slow">dal tronco in rami per fronde, si esprime</prosody>
- 9.<prosody contour="(5%, +81%) (28%, +44%) (42%, -19%) (70%, +41%) (71%, -39%" rate="slow">tutte al ciel tese, con raccolte cime:</prosody>
- 10.<prosody contour="(4%, +43%) (7%, -23%) (32%, -21%) (64%, -39%) (89%, -64%" rate="slow">fermo, rimane il tronco del mistero,</prosody>
- 11.<prosody contour="(22%, +35%) (45%, -10%) (74%, -63%) (94%, +41%) (97%, +72%" rate="slow">e il tronco <emphasis level="strong">s'inabissa</emphasis><break time="50ms" /> ovè più vero.</prosody>

Se lasciamo da parte le due parole aumentate con la marca di “enfasi”, “Vibra” e “s'inabissa” in quanto parole rare e allo stesso tempo indicatori di strutture sintattiche non-canoniche – la prima -, ci sono pause inserite sulla base dell'analisi sintattica, la prima con l'aggiunta di una virgola proprio dopo “Vibra” e la seconda dopo “fermo”. Una pausa di tipo “break” è stata inserita tra due aggiunti, “nell'ansia del pensiero” e “in tutte le sue doglie”; e un'altra tra due clausole alla fine della poesia, “s'inabissa” e “ov'è più vero”. Resta da considerare il movimento del contorno intonativo che come detto precedentemente risponde all'analisi semantica in due marche “foreground” e “background” prodotta dall'analisi aspettuale e da quella del complesso verbale, con la eventuale aggiunta di marcatori del discorso se presenti. Quindi la clausola

governata da “spasima” vede una discesa del contorno, mentre quella governata da “si esprime” una salita, e lo stesso vale per “raccolte” e “rimane”. Per valutare il ruolo della rima in relazione al ritmo complessivo, se cioè le parole di fine verso che contengono la sillaba in rima debbano essere trattate dal ruolo che riveste la metrica o piuttosto dal senso si considera la presenza di scavalcamenti. Nel caso affermativo si dà la prevalenza al secondo metodo in quanto l'inserimento di una pausa a fine verso porterebbe come conseguenza anche una caduta del contorno intonativo e inficerebbe la comprensione del significato. Una semplice giuntura viene inserita automaticamente dal sistema semplicemente staccando le marche prosodiche e questo viene fatto ogni qualvolta non ci sia la presenza di scavalcamenti. Tra i materiali supplementari abbiamo aggiunto due file .wav che contengono la versione di partenza della poesia “Il Pioppo” e quella successiva ottenuta grazie all'analisi di SPARSAR. Allo scopo di mostrare quanto sia importante il contributo di SPARSAR nel caso in cui il senso richiede di superare il confine di verso e anche quello di strofa, includiamo anche le due versioni sempre prodotte da Azure di Microsoft, della poesia “Alla Sera” di Ugo Foscolo, in cui questo fenomeno è ben rappresentato.

## BIBLIOGRAFIA

- [1] Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S., The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) *Prosodic Typology -- The Phonology of Intonation and Phrasing*, (2005).
- [2] Marc Brysbaert, Emmanuel Keuleers and Boris New, Assessing the usefulness of Google Books’ word frequencies for psycholinguistic research on word processing, in *Front. Psychol., Sec. Language Sciences*, volume 2, 2011. downloadable at <https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00027/full>
- [3] Crepaldi, D., Amenta, S., Mandera, P., Keuleers, E., & Brysbaert, M. Frequency estimates from different registers explain different aspects of visual word recognition. *International Meeting of the Psychonomic Society*, (2016).
- [4] Delmonte R., *Cognitive Models of Poetry Reading*, Chapter, in Marcel Danesi (ed.) *Handbook of Cognitive Mathematics*, Springer International Publishing, (2023), pp. 1083-1120.
- [5] Delmonte R., *Poetry and Speech Synthesis*, SPARSAR recites, in *Ricognizioni - Rivista di Lingue, Letterature e Culture Moderne*, II, VI, (2019), pp. 75-95.
- [6] Delmonte, R., *Syntax and Semantics of Italian Poetry in the First Half of the 20th Century*, in *Umanistica Digitale*, n.2, (2018), pp. 35-66.
- [7] Delmonte R., *Expressivity in TTS from Semantics and Pragmatics*, in Vayra, M., Avesani, C. & Tamburini F. (Eds.) *Il farsi e disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio*, Milano: AISV, (2015), pp. 407-427.
- [8] Rodolfo Delmonte, Ciprian Bacalu, SPARSAR: a System for Poetry Automatic Rhythm and Style AnalyzeR, SLATE (2013).
- [9] Silverman, Kim; Beckman, Mary; Pitrelli, John; Ostendorf, Mari; Wightman, Colin; Price, Patti; Pierrehumbert, Janet; Hirschberg, Julia, "TOBI: A Standard for Labeling English Prosody". *International Conference Spoken Language Processing*. Banff, Canada: (1992), 867–870.
- [10] Reddy, Sravana & John Goldsmith, 2010. An MDL-based approach to extracting subword units for grapheme-to-phoneme conversion, in *Proc. HLT-NAACL*, 713-716.
- [11] Genzel Dmitriy, J. Uszkoreit, and F. Och. 2010. “Poetic” statistical machine translation: Rhyme and meter. In *Proceedings of EMNLP*.
- [12] Byrd Roy J. and M. S. Chodorow. 1985. Using an online dictionary to find rhyming words and pronunciations for unknown words. In *Proceedings of the 23rd Annual Meeting of ACL*, 277–283.
- [13] Sonderegger, Morgan. 2011. Applications of graph theory to an English rhyming corpus. *Computer Speech and Language*, 25:655–678.
- [14] Greene E., T. Bodrumlu, K. Knight. 2010. Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation, in *Proceedings of the 2010 Conference on EMNLP*, 524–533.
- [15] Tsur Reuven. 2012. *Poetic Rhythm: Structure and Performance: An Empirical Study in Cognitive Poetics*, Sussex Academic Press, 472.
- [16] Gervas, P. (2001). An expert system for the composition of formal spanish poetry. *Knowledge-Based Systems*, 14(3):181–188.
- [17] Gervas, P. (2010). Engineering linguistic creativity: Bird flight and jet planes. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 23–30.
- [18] Toivanen, J. M., Toivonen, H., Valitutti, A. & Gross, O. 2012. Corpus-based generation of content and form in poetry. In *International Conference on Computational Creativity*, 175–179.
- [19] Agirrezabal Manex, Bertol Arrieta, Aitzol Astigarraga, Mans Hulden, 2013. POS-tag based poetry generation with WordNet, *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 162–166.
- [20] Manurung Hisar Maruli, G. Ritchie, and H. Thompson. 2000a. Towards a computational model of poetry generation. In *Proceedings of AISB Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, 17-20.



- [21] Manurung M.H., G. Ritchie, H. Thompson. 2000b. A Flexible Integrated Architecture For Generating Poetic Texts. in Proceedings of the Fourth Symposium on Natural Language Processing (SNLP 2000), Chiang Mai, Thailand, 7-22.
- [22] Reddy, Sravana & Kevin Knight. 2011. Unsupervised Discovery of Rhyme Schemes, in Proceedings of the 49th Annual Meeting of ACL: shortpapers, 77-82.
- [23] Baayen R. H., R. Piepenbrock, and L. Gulikers. 1995. The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium.
- [24] Hartman, C. (2005). The Scandroid Manual. Online + Hartman, C. (2004). Charles Hartman Programs. Online
- [25] Heuser, R. (2015). Stanford Literary Lab Github Account. Online
- [26] Algee-Hewitt, M., Heuser, R. Kraxenberger, M., Porter, J., Sensenbaugh, J., and Tackett, J. (2014). The Stanford Literary Lab Transhistorical Poetry Project Phase II: Metrical Form. Proceedings, Stanford University, Lausanne.
- [27] Plamondon M. R.: Virtual verse analysis: Analysing patterns in poetry. *Literary and Linguistic Computing* 21, suppl 1 (2006), 127–141.
- [28] Manish Chaturvedi, Gerald Gannod, Laura Mandell, Helen Armstrong, Eric Hodgson. 2012. Myopia: A Visualization Tool in Support of Close Reading. *Digital Humanities* 2012.
- [29] M. Agirrezabal, B. Arrieta, A. Astigarraga, M. Hulden. 2013. “ZeuScansion: a tool for scansion of English poetry,” Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing.
- [30] Nina McCurdy, Vivek Srikumar, Miriah Meyer, 2015. RhymeDesign: A Tool for Analyzing Sonic Devices in Poetry, *Computational Linguistics for Literature*, ACL, Denver, 12-22.
- [31] Manish Chaturvedi, Gerald Gannod, Laura Mandell, Helen Armstrong, Eric Hodgson. 2012. Rhyme’s Challenge: Hip Hop, Poetry, and Contemporary Rhyming Culture. Oxford University Press, *Literary Criticism*.
- [32] Karteek Addanki and Dekai Wu. 2013. Unsupervised Rhyme Scheme Identification in Hip Hop Lyrics using Hidden Markov Models. Proceedings of the 1st International Conference on Statistical Language and Speech Processing (SLSP - 2013), Tarragona, Spain.
- [33] Hussein Hirjee and Daniel Brown. 2009. Automatic Detection of Internal and Imperfect Rhymes in Rap Lyrics. In Proceedings of the 10th International Society for Music Information Retrieval Conference. pages 711-716.
- [34] Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. In Proceedings of Human Language Technologies: NAACL ’09. ACL, Stroudsburg, PA, USA, 308316.
- [35] Alfie Abdul-Rahman, Julie Lein, Katharine Coles, Eamonn Maguire, Miriah Meyer, and Martin Wynne, Chris Johnson, Anne E. Trefethen, Min Chen. 2013. Rule-based Visual Mappings - with a Case Study on Poetry Visualization. In *Computer Graphics Forum*, 32(3):381-390.
- [36] Keim D. A., Oelke D.: Literature fingerprinting: A new method for visual literary analysis. In *IEEE VAST (2007)*, pp. 115–122.
- [37] Oelke D., Bak P., Keim D., Last M., Danon G.: Visual evaluation of text features for document summarization and analysis. In *IEEE VAST (Oct. 2008)*, pp. 75 –82.
- [38] Wattenberg M., Viégas F. B.: The Word Tree, an interactive visual concordance. *IEEE Trans. Visualization & Comp. Graphics* 14, 6 (Nov. 2008), 1221–1228.
- [39] Kao Justine and Dan Jurafsky. 2012. "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry". in Proc. NAACL Workshop on Computational Linguistics for Literature.
- [40] Kaplan, D. (2006). Computational analysis and visualized comparison of style in American poetry. Unpublished undergraduate thesis.
- [41] Kaplan, D., & Blei, D. (2007). A computational approach to style in American poetry. In *IEEE Conference on Data Mining*.
- [42] Hayward, M. (1991). A connectionist model of poetic meter. *Poetics*, 20, 303-317.
- [43] Hayward, M. (1996). Application of a connectionist model of poetic meter to problems in generative metrics. *Research in Humanities Computing* 4. (pp. 185-192). Oxford: Clarendon Press.



## APPENDICE

### Il pioppo di Clemente Rèbora.

Vibra nel vento con tutte le sue foglie  
il pioppo severo;  
spasima l'anima in tutte le sue doglie  
nell'ansia del pensiero:

dal tronco in rami per fronde si esprime  
tutte al ciel tese con raccolte cime:  
fermo rimane il tronco del mistero,  
e il tronco s'inabissa ov'è più vero.

### Alla sera di Ugo Foscolo

Forse perché della fatal quiete  
tu sei l'immagine a me sì cara vieni  
o sera! E quando ti corteggian liete  
le nubi estive e i zeffiri sereni,

e quando dal nevososo aere inquiete  
tenebre e lunghe all'universo meni  
sempre scendi invocata, e le segrete  
vie del mio cor soavemente tieni.

Vagar mi fai co' miei pensieri su l'orme  
che vanno al nulla eterno; e intanto fugge  
questo reo tempo, e van con lui le torme

delle cure onde meco egli si strugge;  
e mentre io guardo la tua pace, dorme  
quello spirito guerrier ch'entro mi rugge.

# Unsupervised Information Extraction from Medieval Charters

Tamás Kovács<sup>1</sup>, Franziska Decker<sup>2</sup>, Daniel Luger<sup>3</sup>,  
Florian Atzenhofer-Baumgartner<sup>4</sup>, Georg Vogeler<sup>5</sup>

<sup>1</sup> Centre for Information Modelling, University of Graz, Austria – [tamas.kovacs@uni-graz.at](mailto:tamas.kovacs@uni-graz.at)

<sup>2</sup> Centre for Information Modelling, University of Graz, Austria – [franziska.decker@uni-graz.at](mailto:franziska.decker@uni-graz.at)

<sup>3</sup> Centre for Information Modelling, University of Graz, Austria – [daniel.luger@uni-graz.at](mailto:daniel.luger@uni-graz.at)

<sup>4</sup> Centre for Information Modelling, University of Graz, Austria – [florian.atzenhofer-baumgartner@uni-graz.at](mailto:florian.atzenhofer-baumgartner@uni-graz.at)

<sup>5</sup> Centre for Information Modelling, University of Graz, Austria – [georg.vogeler@uni-graz.at](mailto:georg.vogeler@uni-graz.at)

## ABSTRACT

Monasterium.net<sup>1</sup> provides a dataset of more than 600,000 European charters. To use this dataset for diplomatics questions on a European scale instead of focusing on individual, regional or collection-specific studies, automatic methods are necessary. In the ERC project “From digital to distant diplomatics”, we combine traditional approaches of diplomatics with state-of-the-art machine learning methods from the fields of natural language processing (NLP). NLP techniques have been largely exploited to solve several tasks, such as content classification. We propose a method for unsupervised information extraction using a Siamese network and few-shot classification on charter digests for various tasks such as categorizing text documents into predefined categories, identifying named entities in a text based on a limited number of labeled examples (NER), and determining whether one text can be inferred from another text.

## KEYWORDS

Diplomatics; Natural Language Processing; Information Extraction; Few-Shot Classification

## 1. INTRODUCTION

The field of diplomatics has always been confronted with a mass of documents that is impossible to study using traditional, manual methods. Conventionally, this problem was tackled by researchers by either reducing their sources to smaller, regional corpora or specific collections, or comparing the results of such studies. Yet, charter collections like Monasterium.net, the largest repository of digitized medieval and early modern charters from all over Europe (majorly from Germany, Italy, Austria, Slovakia, the Czech Republic and Hungary) with a dataset of more than 600,000 charters, cannot be studied as a total. Modern computational methods from the fields of NLP and computer vision offer the chance to transcend regional micro-studies. The ERC project “From Digital to Distant Diplomatics”<sup>2</sup> asks how we can test if insights made on the local, regional, or chancery level can be transferred to overwhelming quantities of charters, especially with regard to charters from the 14th and 15th centuries, which have seen only a few individual studies.

The charter data in Monasterium.net comes in different forms, e.g., as digitized images, abstracts, full texts, information of preserved text witnesses, art historical descriptions and information about existing diplomatic analysis of the object, although only some of this information is available for each document. In the proposed paper, we want to discuss the task of computational information extraction from the charter abstracts (digests), called *regesta* in diplomatics, whose main function is the easy access to core historical information.

## 2. STATE OF THE ART

Information extraction is a task well known in NLP. Several supervised algorithms that learn from labeled or annotated instances have been implemented (e.g., for names, titles, dates or values). Despite the fact that we lack large benchmark research, several publications concentrate on extraction applications in confined and restricted domains, and charters are a good example for this.<sup>3</sup> Still, insufficient annotated examples are a significant impediment. A multiplicity of semantic

---

<sup>1</sup> <https://www.monasterium.net>

<sup>2</sup> <https://www.didip.eu>

<sup>3</sup> E.g., Torres Aguilar and Stutzmann, “Named Entity Recognition for French Medieval Charters”; Boroş et al., “A Comparison of Sequential and Combined Approaches for Named Entity Recognition in a Corpus of Handwritten Medieval Charters”; Bizon Monroc et al., “A Comprehensive Study of Open-Source Libraries for Named Entity Recognition on Handwritten Historical Documents”; Chaffenet, “Noms de Personnes et Noms de Lieux. L’annotation Des Entités Nommées Dans Les Cartulaires Français”; Chastang, Torres Aguilar, and Tannier, “A Named Entity Recognition Model for Medieval Latin Charters”; Torres Aguilar, “La reconnaissance des entités

classifications can be ascribed to the linguistic surface expressions of texts during historical research. For numerous classes, the language variance of expressing the same class is fairly considerable, but the set of elements or their combinations that trigger a semantic pattern in an example is restricted. The manual labeling of sufficient training documents to construct an accurate classifier is frequently prohibitively expensive, especially when sources or the size of Monasterium.net are considered.

In text classification<sup>4</sup>, however, when just a few categories per document text are allocated and where multiple characteristics (i.e., the words in a text) frequently individually trigger the category pattern (e.g., the main verb in an abstract), this may be less of an issue. In addition, even when training an information extraction system in a limited domain, the speed and cost of annotation are significant factors in the development of extraction systems. On the other hand, it is inexpensive to gather a huge amount of unlabeled textual data.

### 3. METHOD

Our proposed approach involves using few-shot classification to extract information from medieval charters. We demonstrate how to train on a small annotated corpus<sup>5</sup> and improve the quality of the learnt classification patterns by utilizing unlabeled data. Our technique turns clustering algorithms, which are unsupervised technologies that rely on unlabeled data, on their heads. This means, it detects the organizing of comparable patterns into comprehensible clusters or groups, enabling the discovery of similarities and differences between patterns and the extraction of relevant inferences from the clusters or, in our case from the charters.

We use a type of deep learning architecture known as a Siamese network. This architecture is composed of two identical neural networks, which are trained simultaneously.<sup>6</sup> The outputs of the two sub-networks are then compared to determine their similarity or dissimilarity, using a similarity metric such as cosine similarity or Euclidean distance. The networks are trained to minimize the distance between the encoded input data and the encoded labels, while maximizing the distance between the encoded input data and the encoded non-labels.<sup>7</sup> This results in the network learning to recognize the patterns in the input data and categorize it into the correct class. The Siamese network architecture is based on the idea that the two sub-networks should share the same weights and biases, so that they can learn a similar representation of the two different inputs they are given. This allows the network to learn to recognize patterns that are common to both inputs, while ignoring patterns that are specific to each input. The shared weights and biases also help to reduce the number of parameters that need to be learned, making the network more efficient.

One of the key advantages of the Siamese network architecture is its ability to learn from a small number of labeled examples. This makes it particularly useful in applications such as few-shot learning, where the goal is to learn a new task with only a few labeled examples.<sup>8</sup> By sharing weights and biases between the two sub-networks, the Siamese network can learn a representation of the input that is robust to small variations in the input.

Our approach relies on a pre-trained language model, which has learned underlying patterns in text from a large corpus, such as Monasterium.net. This pre-trained model can then be fine-tuned with a small number of annotated examples from a new dataset to perform specific tasks, such as named entity recognition, classification, or sentiment analysis. By leveraging the knowledge acquired during pre-training, the model can achieve promising results in these tasks, even when provided with limited labeled data.

It involves three main steps:

1. Pre-processing: The first step involves pre-processing the charter digests to remove any noise, such as optical character recognition (OCR) errors, and standardize the language used. This step is essential to ensure the accuracy of the results obtained from few-shot classification.
2. Pre-training of a language model (optional): In this step, a language model is pre-trained on a large corpus of text, which helps it learn the underlying patterns and structures in the language. This pre-trained model serves as a basis for further fine-tuning on specific tasks. For this purpose, we resort to a bidirectional encoder

---

nommées dans les bases numériques de chartes médiévales en latin”; “Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and BERT-Based Models”; Torres Aguilar, Tannier, and Chastang, “Named Entity Recognition Applied on a Data Base of Medieval Latin Charters. The Case of Chartae Burgundiae”; Karsvall and Borin, “SDHK Meets NER: Linking Place Names with Medieval Charters and Historical Maps”.

<sup>4</sup> Kowsari et al., “Text Classification Algorithms.”

<sup>5</sup> Work done in the context of our research project.

<sup>6</sup> Koch, Zemel, and Salakhutdinov, “Siamese Neural Networks for One-Shot Image Recognition”.

<sup>7</sup> Mueller and Thyagarajan, “Siamese Recurrent Architectures for Learning Sentence Similarity.”

<sup>8</sup> Vinyals et al., “Matching Networks for One Shot Learning.”

representations from transformers (BERT) based language model that we have already trained on the Monasterium dataset.

3. Generating limited support set: The final step involves fine-tuning the pre-trained model on a small set of annotated charters ( $n = 5, 10, 15 \dots 100$ ), which act as the few-shot examples. The annotations can be created manually or automatically using existing techniques, such as rule-based systems.

During the training phase, the model is optimized to recognize particular data types, such as names, dates, languages, and locations. Typically, the process of fine-tuning entails trial-and-error to ensure that the model produces high-quality results. Manual error correction and iterative improvement are options for enhancing the efficacy of the model.

In this specific approach, a prototypical network is utilized. The aim is to train the model to classify new examples based on a small number of labeled examples, referred to as the support set. The code includes a class called Embedding Extractor that uses a long short-term memory (LSTM) neural network to extract an embedding from a sequence of input vectors. The prototypical network class is then defined, which accepts a set of support embeddings as input and calculates the prototypes for each class. The network computes class prototypes as the mean embeddings of the support set instances and classifies new instances based on their similarity to these prototypes.

Throughout the training procedure, the Embedding Extractor model is updated and the loss is computed using the Cross Entropy Loss criterion. The evaluate function computes the model's accuracy on the test set and provides an evaluation of its performance.

## 4. RESULTS

The proposed approach is expected to provide a cost-effective and accurate method for extracting information from medieval charters. The use of few-shot classification will reduce the need for a large annotated dataset and enable the model to identify patterns in the text that can be used to extract information. The strategy will be assessed using a limited but meticulously chosen group of medieval charters, and the level of accuracy achieved will be compared to that of conventional supervised learning methods.

Our evaluation set for different NLP tasks is called "1000 NLP Charters", containing 1,000 diverse charters which represent the Monasterium data in a balanced way also in our research timeline (13th to 16th century). The dataset covers various aspects of the task and includes examples from different domains, styles, and difficulty levels to ensure a comprehensive assessment of the model's performance. It is representative of real-world scenarios, reflecting the distribution of the data the model is expected to encounter in practical applications. Consequently, it also contains OCR errors and remains of editorial tags as well as non-charter texts like remnants of critical apparatuses or charter abstracts instead of the charter tenors. The "1000 NLP Charters" evaluation set is distinct from the training data to prevent overfitting and ensure an unbiased evaluation of the model's generalization capabilities. While the "1000 NLP Charters" dataset is still under development, with ongoing manual annotation efforts focusing on key tasks such as language detection, extended Named Entity Recognition, formulaic language detection, OCR error identification, and similarity evaluation, it can already be utilized for specific evaluation purposes in its current state. On entity extraction the proposed approach obtained 94,5% F1 score ( $n=50$ ).

Although the current result is below the task specific state of the art, it is worth to stress again that few-shot classification can enable researchers to extract valuable information from these documents with minimal annotation, reducing the time and resources required for manual annotation. This approach can be used to automatically annotate a larger dataset by extracting for instance named entities from unannotated text. This newly generated labeled dataset can be used to train or fine-tune other task-specific models or approaches. Furthermore, it can be extended to other historical documents, making it a valuable tool for the extraction of historical information from large datasets.

## 5. CONCLUSIONS

Medieval charters are essential historical documents that contain vast amounts of information about the social, economic, and political life of the past. Our work is based on the data found in Monasterium.net with more than 600,000 charter entries. In the proposed poster, we discuss the challenges and promises of automated information extraction from the *regesta* texts, one of the charter data appearances in Monasterium.net. We propose to employ the NLP method of few-shot classification using a Siamese neural network as a way to reduce the time and resources needed to train models, taking advantage of the technologies' ability to perform with unlabeled data. The successful automated extraction of information from unlabeled charter abstracts enables diplomatics to conduct large-scale investigations on a transnational level, which previously could not be accomplished by manual efforts, and opens the doors for the progression to distant diplomatics.

## 6. ACKNOWLEDGEMENTS

All of the contributing authors are part of the ERC project “From Digital to Distant Diplomatics”, located at the Centre for Information Modelling of the University of Graz. Tamás Kovács has developed the concept and the code of the presented approach. Franziska Decker and Daniel Luger contributed to the work with data evaluation and diplomatics research. Florian Atzenhofer-Baumgartner collected the dataset for the experiments, whereas Georg Vogeler supported the research with comments and his supervision. Our team would like to thank the European Research Council for their generous funding, as well as the International Centre for Archival Research (ICARus) and the involved archives for providing their data on Monasterium.net.

## REFERENCES

- [1] Bizon Monroc, Claire, Blanche Miret, Marie-Laurence Bonhomme, and Christopher Kermorvant. “A Comprehensive Study of Open-Source Libraries for Named Entity Recognition on Handwritten Historical Documents.” In Document Analysis Systems 2022. Proceedings of the 15th IAPR International Workshop, La Rochelle, France, May 22–25, 2022, edited by Seiichi Uchida, Elisa Barney, and Véronique Eglin, 429–44. Lecture Notes in Computer Science 13237. Cham: Springer International Publishing, 2022. [https://doi.org/10.1007/978-3-031-06555-2\\_29](https://doi.org/10.1007/978-3-031-06555-2_29)
- [2] Boroş, Emanuela, Romero, Verónica, Maarand, Martin, Zenklová, Kateřina, Křečková, Jitka, Vidal, Enrique, Stutzmann, Dominique, and Kermorvant Christopher. “A Comparison of Sequential and Combined Approaches for Named Entity Recognition in a Corpus of Handwritten Medieval Charters.” In Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition (ICFHR 2020), 79–84. The Institute of Electrical and Electronics Engineers, 2020. <https://doi.org/10.1109/ICFHR2020.00025>
- [3] Chaffenet, Paul. “Noms de Personnes et Noms de Lieux. L’annotation Des Entités Nommées Dans Les Cartulaires Français.” Himanis (blog), 10 February 2021. <https://himanis.hypotheses.org/1059>
- [4] Chastang, Pierre, Sergio Torres Aguilar, and Xavier Tannier. “A Named Entity Recognition Model for Medieval Latin Charters.” Digital Humanities Quarterly 15, no. 4 (2021). <http://www.digitalhumanities.org/dhq/vol/15/4/000574/000574.html>
- [5] Karsvall, Olof, and Borin, Lars. “SDHK Meets NER: Linking Place Names with Medieval Charters and Historical Maps.” In Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), 38–50. CEUR Workshop Proceedings 2084, 2018. <http://ceur-ws.org/Vol-2084/>
- [6] Koch, Gregory, Zemel, Richard, and Salakhutdinov, Ruslan. “Siamese Neural Networks for One-Shot Image Recognition.” In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), 2:1920–28. Lille, 2015. <https://api.semanticscholar.org/CorpusID:13874643>
- [7] Kowsari, Kamran, Jafari Meimandi, Kiana, Heidarysafa, Mojtaba, Mendu, Sanjana, Barnes, Laura, and Brown, Donald. “Text Classification Algorithms: A Survey.” Machine Learning on Scientific Data and Information, Information Applications, 10, no. 4 (April 2019). <https://doi.org/10.3390/info10040150>
- [8] Mueller, Jonas, and Thyagarajan, Aditya. “Siamese Recurrent Architectures for Learning Sentence Similarity.” In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Vol. 1. Phoenix: AAAI Press, 2016. <https://doi.org/10.1609/aaai.v30i1.10350>
- [9] Torres Aguilar, Sergio, Tannier, Xavier, and Chastang, Pierre. “Named Entity Recognition Applied on a Data Base of Medieval Latin Charters. The Case of Chartae Burgundiae.” In HistoInformatics 2016. The 3rd HistoInformatics Workshop. Proceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016). Co-Located with Digital Humanities 2016 Conference (DH 2016). Krakow, Poland, July 11, 2016, 67–71. CEUR Workshop Proceedings 1632, 2016. <https://ceur-ws.org/Vol-1632/>
- [10] Torres Aguilar, Sergio, and Stutzmann, Dominique. “Named Entity Recognition for French Medieval Charters.” In Proceedings of the Workshop on Natural Language Processing for Digital Humanities, 37–46. Silchar: NLP Association of India (NLP AI), 2021. <https://hal.archives-ouvertes.fr/hal-03503055>
- [11] Torres Aguilar, Sergio. “La reconnaissance des entités nommées dans les bases numériques de chartes médiévales en latin: Le cas du Corpus Burgundiae Medii Aevi (Xe–XIIIe siècle).” Médiévales. Langues, Textes, Histoire, no. 73 (2017): 47–65. <https://doi.org/10.4000/medievales.8182>
- [12] Torres Aguilar, Sergio. “Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and BERT-Based Models.” In Proceedings of the LREC 2022 Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), 119–28. Marseille: European Language Resources Association, 2022. <https://hal.archives-ouvertes.fr/hal-03703239>
- [13] Vinyals, Oriol, Blundell, Charles, Lillicrap, Timothy, Kavukcuoglu, Koray, and Wierstra, Daan. “Matching Networks for One Shot Learning.” In Advances in Neural Information Processing Systems 29 (NIPS 2016). Proceedings, 29:3630–38. Barcelona: Curran Associates, Inc., 2016. <http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning>

# The challenges of sense granularity in word sense disambiguation for Latin

Barbara McGillivray<sup>1</sup>, Iacopo Ghinassi<sup>2</sup>, Paola Marongiu<sup>3</sup>

<sup>1</sup> King's College London, United Kingdom – barbara.mcgillivray@kcl.ac.uk

<sup>2</sup> Queen Mary University of London, United Kingdom – i.ghinassi@qmul.ac.uk

<sup>3</sup> Université de Neuchâtel, Switzerland – paola.marongiu@unine.ch

## ABSTRACT

Developing algorithms for word sense disambiguation (WSD) for historical languages can enable large-scale analyses of diachronic semantics, with beneficial impacts reaching far beyond historical linguistics. Latin is in a particularly favourable position among historical languages given the number of digital language resources and tools available for it. Building and expanding on the state-of-the-art research in Latin WSD, this poster presents preliminary results from a WSD algorithm based on BERT embeddings trained from dictionary quotations from the Lewis & Short dictionary. Our work furthers the state of the art by assigning the most likely sense of a lemmas among all its possible macro-senses and we discuss the challenges of dictionary sense granularity in Latin WSD.

## KEYWORDS

Word sense disambiguation; Latin corpora; computational semantics; historical semantics

## 1. STATE OF THE ART AND OBJECTIVES

A wealth of Latin textual resources in digital form are available and can now be linked at the level of their lemmas [8]. What is missing is the possibility to connect texts and lexical resources at the level of word senses, which would enable us to conduct corpus analyses of the distributions of words' meaning. Beyond historical semantics research, this is relevant to digital humanities more broadly, allowing for large-scale analyses of word meaning in fields such as history, literary studies, and classics. Word Sense Disambiguation (WSD) associates word instances in a corpus to their most appropriate sense and can greatly help annotate corpora for semantic analyses. [1], [2], and [4] have developed WSD for Latin, training classification systems on dictionary quotations mapped to dictionary senses using BERT embeddings [3]. However, such previous studies have dealt with only the first two macro-senses of lemmas; this excludes a high number of lemmas, since only 10,663 (27%) out of 40,023 entries in the Lewis and Short dictionary (L&S) [5] have two macro-senses, as shown in Figure 1. For instance, *sentio* has three macro-senses in L&S: 'to feel, perceive with senses'; 'to notice, perceive, observe'; 'to think, suppose, to give one's opinion, to decide'. Moreover, important nuances are only captured at more granular levels than macro-senses. For instance, the lemma *spiritus* has two macro-senses in L&S: (1) 'air, breathing' and (2) 'spirit, energy, soul'. Sense (1), which is concrete, also comprises the abstract sub-senses 'breath of life, life' and 'divine inspiration'. This poster presents the preliminary results on an algorithm for Latin WSD trained all macro-senses in the L&S dictionary and we analyse to what extent sense granularity affects the algorithm's performance.

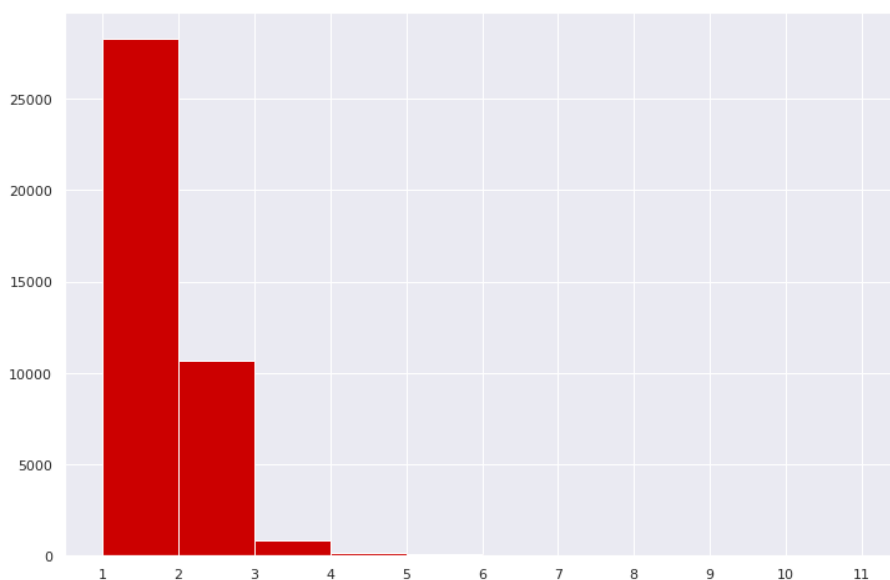


Figure 1. Number of L&J entries by number of macro-senses. Most entries only have one or two macro-senses, while very few have three, four or five, and even fewer have more than five.

## 2. DATA AND METHODS

We trained state-of-the-art BERT embedding models developed for Latin [2]<sup>1</sup>, available at <https://github.com/dbamman/latin-bert>. Focusing on 134 nouns, adjectives, and verbs for which at least 10 quotations were available, we mined the L&S entries to extract all their macro-senses. We performed two main experiments: first, we grouped together all quotations associated to L&S macro-senses and their sub-senses and trained the algorithm on all macro-senses; second, we trained the algorithm of all micro-senses indicated by upper-case letters in L&S entries. In the example of the entry for *spiritus*, in the first experiment we considered two main senses and their quotations: first, all the quotations included in macro-sense I (with all its subsenses, A, B, B.1, B.2, B.3, C.1, C.2, C.3, C.4, C.5, and C.6), and second, all quotations included in macro-sense II (with its subsenses A.α, A.β, B.α, B.β, and B.b.2). In the second experiment, we considered four senses: I.A (with its quotations), I.B (with all its quotations and the quotations of its subsenses B.1, B.2, and B.3), I.C (with all the quotations of its subsenses C.1, C.2, C.3, C.4, C.5, and C.6), II.A (with all the quotations of its subsenses A.α and A.β) and II.B (with all the quotations of its subsenses B.α, B.β, and B.b.2). See below for the structure of the entry for the lemma *spiritus* in L&S with the main senses and number of quotations for each sense.

I.a breathing or gentle blowing of air, a breath, breeze

**I.**

**A.** In general (5 quotations)

**B.** In particular

1. The air (4 quotations)

2. An exhalation, smell, odor (4 quotations)

3. Breathed air, a breath (17 quotations)

**C. Transf.**

1. a breathing (4 quotations)

2. The breath of a god, inspiration (2 quotations)

3. The breath of life, life (6 quotations) to expire, die (3 quotations)

4. a sigh (0 quotations)

5. In language, a breathing or aspiration (0 quotations)

6. The hiss of a snake (0 quotations)

**II. Trop.**

**A.** A haughty spirit, haughtiness, pride, arrogance; also, spirit, high spirit, energy, courage (esp. freq. in the plur.; syn. animi).

<sup>1</sup> We used BERT-base-cased from the Huggingface python library (<https://github.com/dbamman/latin-bert>) as the starting model for training.



- ( $\alpha$ ). Singular form (11 quotations)
- ( $\beta$ ). Plural form (12 attestations)
- B.** Spirit, soul, mind.
- ( $\alpha$ ). Singular form (10 quotations)
- ( $\beta$ ). Plural form (1 quotation)
- b.** a beloved object (0 quotations)
- 2.** a spirit (1 quotation); the Holy Ghost, Holy Spirit (2 quotations); an evil spirit (1 quotation)

In the first experiment, we fine-tuned BERT for each lemma and trained a multi-class classifier on the L&S quotations to predict the most likely sense among up to five of the lemma’s macro-senses, the highest number in L&S. In the second experiment, we performed the same operations but following the senses as explained above. We performed inference on the LatinISE corpus [6] for quantitative and comparative analyses. Our code is available at <https://github.com/Ighina/LatinISE-WSD/tree/master/wsd>.

### 3. RESULTS AND DISCUSSION

Figure 1 shows that the first two macro-senses are over-represented in the L&S dictionary, while the number of quotations available for the third macro-sense and the following ones is very low, leading to few predictions for these senses in a separate inference corpus such as LatinISE. Figure 2 shows the distribution of predicted macro-senses in the first experiment on the 134 lemmas in the LatinISE corpus, which can be further employed in quantitative analyses of Latin lexical semantics.

We computed accuracy metrics for the same test set used by [2]: the results of our first experiment reached a weighted average F1 of 70.83%, with an average precision of 75.39%, an average recall of 71.47% and an average accuracy of 72.30%, comparable with the state of the art [2,4]. The results of the second experiment show a drop in performance across all metrics, as displayed in Table 1. Considering all micro-senses leads to suboptimal performance, which is likely a consequence of data scarcity, where very specific senses have fewer examples from which the algorithm can learn.

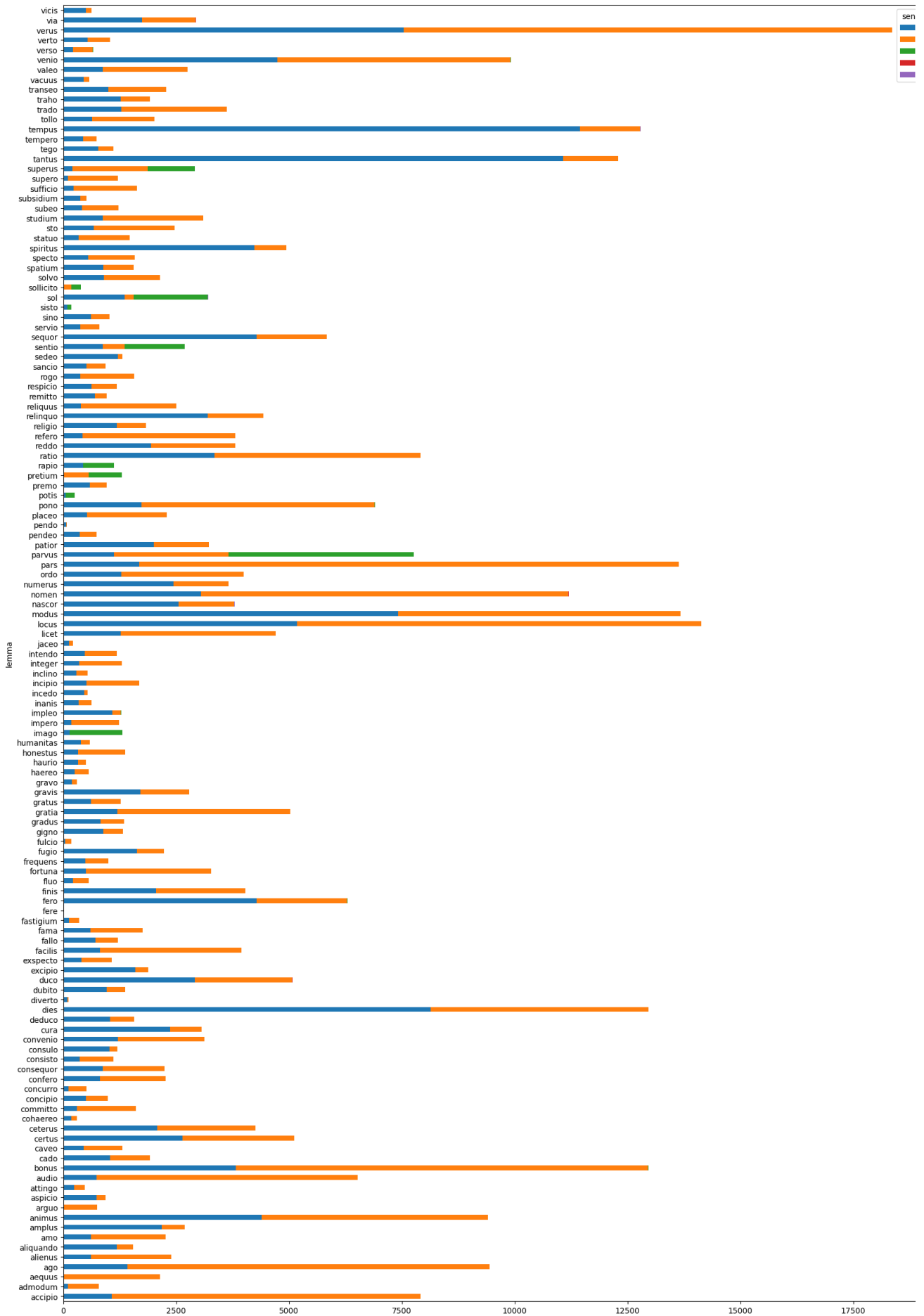


Figure 2. Distribution of predicted sense counts of our WSD algorithm on the LatinISE corpus.

	F1	PRECISION	RECALL	ACCURACY
EXPERIMENT 1	70.83%	75.39%	71.47%	72.30%
EXPERIMENT 2	47.64%	42.60%	56.01%	49.12%

Table 1. Results of our experiments, showing accuracy metrics: average F1, precision, recall and accuracy.

We further evaluated the algorithm on a random selection of ten LatinISE sentences for each macro-sense assigned to the lemmas. We notice that sense granularity has a significant impact on the results. For example, in 80% of the cases of the sample in which *spiritus* was tagged with the second sense ‘spirit, energy, soul’ are correct, but all the occurrences tagged with the concrete meaning ‘air, breathing’ were incorrect. This may be explained by the fact that, while the second macro-sense for *spiritus* only refers to abstract concepts, the first one also includes more abstract sub-senses such as ‘breath of life, life’ or ‘divine inspiration’. This may have influenced the model training by associating sentences with more abstract context with the concrete macro-sense.

The impact of sense granularity on the performance of the algorithm is even more striking when looking at the results of the experiment with micro-senses (experiment 2). An interesting example is given by the results of the automatic annotation for the word *manus*. This lemma has two macro-senses in L&S: the first one is associated with the literal meaning ‘hand’. It does not have any micro-senses associated with it, but it contains all the senses more or less derived from or linked to this literal meaning, such as ‘at hand, near’; ‘closeby’, ‘by hand i.e., artificially’ among others. The second macro-sense collects all abstract senses that can be associated with *manus* and it is specified into 13 micro-senses: ‘the hand as the instrument used in fight’ i.e., ‘personal valor, bravery’ and ‘force, violence, fighting’; ‘the finishing touch’; ‘handwriting’ or ‘workmanship’; ‘a side’; ‘a stake’; ‘a thrust, hit, blow’ (in fencing); ‘the trunk (of an elephant)’; ‘the fore-paws’; ‘the branches’; ‘iron hooks with which an enemy’s ship was grappled’; ‘an armed force’; ‘labor, hands’; ‘power’ and ‘legal power of a husband over his wife’; ‘an arrest’. The difference between the two macro-senses in terms of the number of quotations is glaring. While the first macro-sense ‘hand’, albeit with all its derived senses, is associated with more than 100 quotations, the micro-senses grouped under macro-sense II are associated with a range of one to about 10 quotations each, depending on the micro-sense. Sub-senses are often represented by a much lower number of quotations compared to macro-senses. Therefore, training the algorithm with micro-senses quotations leads to lower performances due to data sparsity problems. In the case of *manus*, the model is trained on a much higher number of quotations for sense I compared to micro-senses II.A, II.B, II.C etc., and ends up overestimating the presence of senses I compared to senses II.A, II.B, II.C, and so on. As a result, out of 8 sentences in the test set containing the word *manus*, the only two sentences in which the word had sense I were both tagged correctly. The others, in which the word expressed senses II.B, II.C, II.D., IIF., and II.A, were all incorrectly tagged with sense I, except for one, which was correctly assigned sense II.A. A strategy to address such data sparsity problems in the future is by complementing the dictionary quotations with sense-annotated corpus data such as [7].

## REFERENCES

- [1] Bamman, David, and Gregory Crane. “Computational Linguistics and Classical Lexicography”. *Digital Humanities Quarterly* 3, no. 1 (2009).
- [2] Bamman, David, and Patrick Burns. “Latin BERT: A Contextual Language Model for Classical Philology”. Arxiv: arXiv:2009.10053 (2020).
- [3] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, 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*, 4171–4186. Association for Computational Linguistics, 2019.
- [4] Lendvai, Piroska, and Claudia Wick. “Finetuning Latin BERT for Word Sense Disambiguation on the Thesaurus Linguae Latinae”. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, Taipei, Taiwan, 37–41, Association for Computational Linguistics, 2022.
- [5] Lewis, Charlton T., and Charles Short. *A Latin Dictionary*. Oxford: Clarendon Press, 1879.
- [6] McGillivray, Barbara, and Adam Kilgarriff. “Tools for historical corpus research, and a corpus of Latin”. In *New Methods in Historical Corpus Linguistics*, edited by Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt. Tübingen: Narr, 2013.
- [7] McGillivray, Barbara, Daria Kondakova, Annie Burman, Francesca Dell’Oro, Helena Bermúdez Sabel, Paola Marongiu, and Manuel Márquez Cruz. “A new corpus annotation framework for Latin diachronic lexical semantics”. *Journal of Latin Linguistics* 21 no. 1 (2022): 47–105. <https://doi.org/10.1515/joll-2022-2007>
- [8] Passarotti, Marco Carlo, and Francesco Mambrini. “Linking Latin: Interoperable Lexical Resources in the LiLa Project”. In *Building new resources for historical linguistics*, 103–124. Pavia: Pavia University Press, 2022. <https://doi.org/10.5281/zenodo.5994271>

# Machine Learning Algorithm for the Scansion of Old Saxon Poetry

Irene Miani<sup>1</sup>, Marina Buzzoni<sup>2</sup>, Gianluca Lebani<sup>3</sup>, Alessandro Torcinovich<sup>4</sup>

<sup>1</sup> Università Ca' Foscari Venezia, Italy – 888260@stud.unive.com

<sup>2</sup> Università Ca' Foscari Venezia, Italy – mbuzzoni@unive.it

<sup>3</sup> Università Ca' Foscari Venezia, Italy – gianluca.lebani@unive.it

<sup>4</sup> Eidgenössische Technische Hochschule Zürich, Switzerland – alessandro.torcinovich@inf.ethz.ch<sup>4</sup>

## ABSTRACT

Several scholars designed tools to perform the automatic scansion of poetry in many languages, but none of these tools deal with Old Saxon or Old English. This project aims to be a first attempt to create a tool for these languages. We implemented a Bidirectional Long Short-Term Memory (BiLSTM) model to perform the automatic scansion of Old Saxon and Old English poems. Since this model uses supervised learning, we manually annotated the *Heliand* manuscript, and we used the resulting corpus as labeled dataset to train the model. The evaluation of the performance of the algorithm reached a 97% for the accuracy and a 99% of weighted average for precision, recall and F1 Score. In addition, we tested the model with some verses from the *Old Saxon Genesis* and some from *The Battle of Brunanburh*, and we observed that the model predicted almost all Old Saxon metrical patterns correctly misclassified the majority of the Old English input verses.

## KEYWORDS

Old Saxon meter; automatic scansion; annotation; machine learning

## 1. INTRODUCTION AND BRIEF STATE OF THE ART

Meter can be considered as a set of rules and principles that define the rhythmic beat of a poem based on a series of common structures. These structures were studied by several scholars among which we can mention Suzuki's work on Old English and Old Saxon meter<sup>[18]</sup>. The analysis of meter is also called scansion; its goal is to mark the rhythmic patterns across all verses of a poem<sup>[1]</sup>. When done manually, this task is very challenging and time-consuming; for years scholars have tried to develop tools to automatize this analysis. Some scholars, such as Logan<sup>[13]</sup>, Beaudoin and Yvon<sup>[3]</sup>, Gervas<sup>[5]</sup>, *Pede Certo*<sup>[15]</sup>, used rule-based approaches. Other scholars, as Hayes<sup>[8]</sup>, used data driven approaches. More recent studies tried to use both supervised and unsupervised machine learning models to ease this task, in particular we can cite the work of Agirrezabal et al.<sup>[2]</sup>, and Tueller<sup>[20]</sup>. The majority of these tools focus on languages such as English, French or German. However, to the extent of our knowledge, there exists no tool for the automatic scansion of Old English or Old Saxon poems. Thus, the goal of this project is to create the first tool able to perform this task in both languages. In this paper, we will discuss firstly the materials and methods. Then, we will present the results of the evaluation of the model and suggestion for future studies.

## 2. MATERIALS AND METHODS

### 2.1. The annotation of the *heliand*

In order to be trained, supervised learning models require a labeled dataset, from which the machine extracts a pattern that will allow to generalize over new data. Despite the presence of several corpora, we were not able to find a metrically annotated corpus suitable for our purposes. Thus, we decided to annotate a poem manually and use it as dataset for the model. Several poems were taken into consideration, but we concluded that the best choice was the *Heliand*<sup>[14]</sup>, an Old Saxon manuscript. We chose this poem, firstly because the almost 6000 lines, of which the poem is composed, can be considered a sufficient number of data to train the model. Secondly, Suzuki performed the annotation of the *Heliand* in 2004<sup>[18]</sup>, and he indicated, for each verse, its metrical type and pattern. Therefore, we relied on his annotation in order to create our corpus. Thirdly, we wanted to test the hypothesis that it was possible to train the algorithm with the Old Saxon syllables also to predict the scansion of Old English verses.

In 1885, Eduard Sievers proposed the first scholarly classification of the Old Germanic meter<sup>[17]</sup>, which is now considered a standard for the scansion of poetry. His proposal was further developed by Bliss<sup>[4]</sup> and Suzuki<sup>[18]</sup>, among many other

scholars. Each study focused on five main metrical types, labeled as Type A, B, C, D, and E; each metrical type was associated to a specific pattern that corresponds to the rhythmic structure of the verse. Within these macro classes, other subcategories are possible, on the basis of the feet structure. As already mentioned, to create the dataset we relied on Suzuki’s metrical scansion of the *Heliand*. The annotation of the verses is TEI conformant [19]. We represented it as a tree-like structure, having the poem at the highest level and the single word at the lowest one. In Old Germanic meter, each line was composed of two verses or half-lines, and their division depended on where the caesura fell. Each verse was made of one or more number of feet. Each word was considered a separate element of the foot. The following diagram illustrates this tree-like structure.

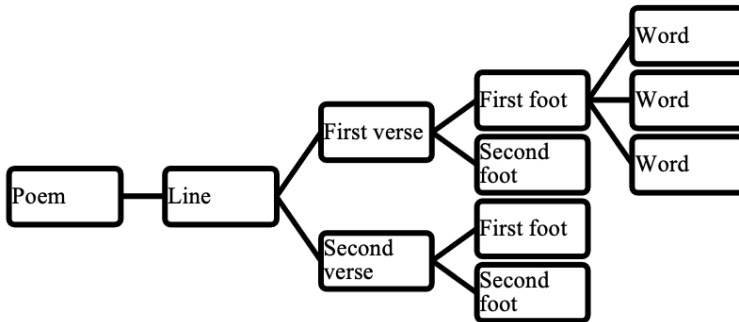


Figure 3. Tree Structure of the Annotation of the *Heliand*.

Starting from this idea, we constructed our annotation of each component according to the TEI guidelines since they are a universally recognized markup language when the goal is to represent the features of a text. Besides the poem and the division in lines, the macro structure is the verse or half-line, which was encoded with the `<|>` element together with the `@n` attribute that represents the number of the line. Then, we used a series of `<seg>` elements to encode all the levels illustrated in the diagram. The first is the `<seg>` element of the verse and each one contains two attributes: `@type` and `@met`. The value of the `@type` attribute corresponds to the metrical type enucleated by Suzuki, while the value of the `@met` attribute corresponds to the metrical pattern. These two attributes are almost always present also in the other segments. The second segment is the one of the feet that will have an additional `@subtype` attribute containing an ordinal number to distinguish the first foot from the second, and eventually, from the other feet. Inside the foot, there can be two types of (sub-)segments: the word segment with a `@met` attribute, and the punctuation segment. Each foot is divided by the `<caesura/>` element. This structure is repeated for all the feet of the verse, until the verse is complete. The only exceptions to this structure are the verses that contain foreign words. These verses were annotated by Suzuki as type F, and he did not specify the metrical pattern for these words. In our annotation, we labeled these verses as type Y. The main differences from the basic structure are the missing `@met` attributes, and the use of the values of the `@subtype` attributes to mark these words are foreign. The following figure represents a sample of our annotation of the first verse of line 1 of the *Heliand*.

```

<| n="1">
  <seg type="A1" met="pxx|Px">
    <seg type="foot" subtype="first" met="pxx">
      <seg type="word" met="pxx">Manega</seg>
    </seg>
    <seg type="foot" subtype="second" met="Px">
      <seg type="word" met="Px">uuâron</seg>
      <seg type="punctuation">,</seg>
    </seg>
  </seg>
</seg>
<caesura/>

```

Figure 2. Example of the Annotation of the *Heliand*.

The metrical labels are another important element to discuss. Suzuki used a series of labels to identify different aspects of the verse; for instance, he used “P” and “p” values to mark respectively long and short primary-stressed syllables; “S” and “s” is used for long and short secondary-stressed syllables, and “X” and “x” for long and short unstressed syllables. We train the model to predict a selection of these labels, because this would have allowed the model to perform a more accurate analysis of the verse focusing only on the main features of the verse necessary to determine its metrical pattern. We considered just three labels: ‘x’ for unstressed syllables, ‘p’ for primary stressed ones and ‘s’ for secondary stressed

syllables. However, we resulted with an imbalance dataset; in fact, our corpus was made of 53.578 syllables of which the 90% belonged to class “x”, the 9% to class “p” and the 1% to class “s”. We tried to solve this problem by using the Focal loss since this function was designed in order to decrease the importance of an example already well classified. Once the dataset was ready, we considered an 80% for the train set and a 20% for the test set.

## 2.2. The machine learning model

The structure of the model is made of a series of passages that lead to the prediction of the metrical pattern of the verses and their metrical type. The following diagram sums up the passages of this model:

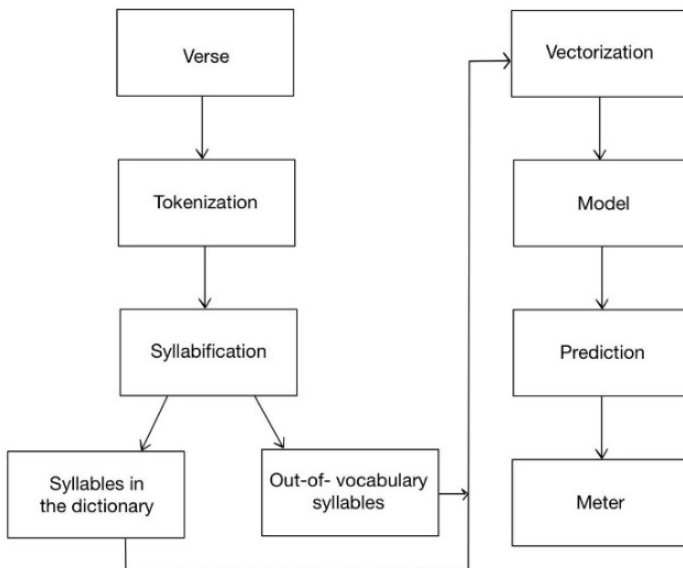


Figure 3. The Structure of the Model.

The inputs are verses already divided into feet by the user; then, these data are processed to make them suitable for the model. After the tokenization, each word is passed on to SonoriPy, a syllabifier created by Estes and Hensch [9]. We used this tool because we could not find a syllabifier specific for Old Saxon or Old English. Furthermore, the user has the possibility to make this tool target specific by adding missing alphabet letters from the target language. In addition, this tool is based on the Sonority Sequencing Principle; thus, each syllable would have a vowel to which we can associate our labels <sup>1</sup>. Once that the syllabification is complete, the syllables are searched in a syllables dictionary, in which each syllable is associated to an index. If the syllable is present in the dictionary, then, it will be replaced by the index in the vector that will be sent as input for the model. If the syllable is not present, it will be considered as an out-of-vocabulary syllable, and the model will simply ignore its presence. Before the vector is sent to the model, it needs to be padded to reach the same length of the verses with which we trained the model.

The BiLSTM model was built using Keras library [11]; we chose this library instead of others because the documentation and the tools are user-friendly. It offers enough support to create efficient models and reduces also the time required for their creation. We implemented a sequential model in which the first layer is the Input layer that instantiates a Keras tensor and is the entry point of the algorithm. These inputs are then sent to an Embedding layer which converts the indexes into a dense vector of fixed size. We set the mask\_zero attribute to True, so that the model will ignore the padded elements. The dense vector is sent to the Bidirectional layer, in which we specified the type of neural network, the LSTM which is based on Hochreiter and Schmidhuber’s proposal of 1997 [10]. In this layer, we specified that the model should return a sequence from the last cell state, instead of a single value. The output of this layer will then pass to a Time Distributed layer connected with a Dense layer; so, the model is able to compute a prediction for each time step in the input sequence. The last layer is the Activation one that applies a softmax function to the output, to convert the vector of values into a distribution of probabilities.

The model is then configured for the training with the compile method in which we specified the optimizer, the loss and the metrics. We used the Adam optimizer with a learning rate of 0.01. For what concerns the loss, since the dataset is

<sup>1</sup> A problem that we faced is the fact that in Old Saxon there is the vowel cluster <uu>, which corresponds to the sound /w/ and not to /u:/; thus, SonoriPy was not able to consider these two letters not as two vowels, so, we decided to replace <uu> with <w>, in order to avoid this problem.

unbalance, we used the `categorical_focal_loss` designed by Griffo [7] which is based on the standard Focal loss function, but it has been modified to deal with multi-classification problems. Griffo added a hyper-parameter, `alpha`, that corresponds to an array of weights for the different classes according to which the model modifies the loss for each class.

The hyper-parameters of the model were computed using Keras Tuner [12], a tool offered by Keras Library to find the best values for these parameters. We used this tool together with the Bayesian Optimization in order to find the best set of values for the model. In our case we searched for the following values: the units of the LSTM, the learning rate of the Adam optimizer and `gamma` of the categorical focal loss. Keras tuner tested 15 possible combinations and concluded that 160 for the units, 0.01 for the learning rate and 2 for `gamma` were the best values to reach the highest accuracy of the model.

The output of the model is, then, sent to a metrical type prediction function which assigns the corresponding metrical type to each metrical pattern. According to the number of metrical patterns, the function will apply different processes. If the metrical patterns are one or more than two, the function will use the FuzzyWuzzy library [16] to extract the most suitable metrical types for the input. The function will base the search on the patterns provided in Suzuki's summary table in which there are the main five metrical types and their subcategories. Each input pattern will be compared with the patterns provided by Suzuki and the function will find the five most similar ones. Each of the five candidates will be associated to one of the main five metrical types, and since the subcategories are also considered in the confrontation process; to avoid the repetition of a metrical types, we used a Python function, the `set`, to eliminate duplicates. If the input is made of only two metrical patterns, then, the function will use the regular expressions to search for the correct metrical type. Only when the metrical patterns are two, will the function return "No match found" if it is not able to associate a metrical type to the pattern.

### 3. RESULTS

The model reached an accuracy of the 97% which is quite satisfying for a first approach in this field; however, since the dataset is not balanced, this metric alone is not sufficient to conclude that this model performs well. Thus, we retrieved the plots of loss and validation loss which shown that the loss curve decreases gently in both train and test, and the curves overlap, so the model is correctly learning the pattern. We considered also the plot of accuracy and validation accuracy, in which we identified a little bit of overfitting. We computed also a confusion matrix from which we extracted Precision, Recall and F1 Score for each label. For what concerns the Precision, the results were high for each label: labels 'p' and 'x' reached the 98%, while 's' the 95%. 'p' and 'x' showed higher values also for Recall reaching respectively the 95% and the 99%; while 's', due to its lower frequency in the dataset, reached the 67%. A 96% value for the F1 Score of 'p' was reached, and for 'x', it reached the 99%. The lower F1 Score was the one of 's' which corresponded to the 78%. The weighted average for precision, recall and F1 score reached the 99% for each metric, which allowed us to conclude that despite the imbalance dataset, the model is able to predict correctly the metrical pattern of the verses.

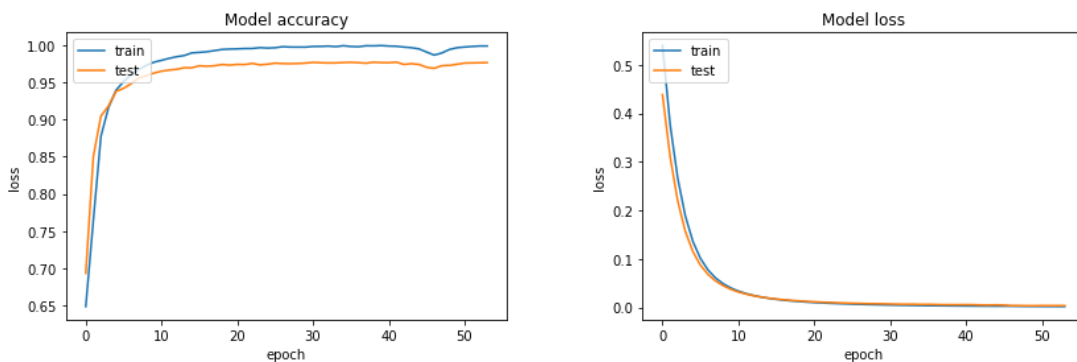


Figure 4. The Plots of Accuracy and Loss Function.

We decided to test our model also with different poems. We first tested it on some verses from the *Old Saxon Genesis* [6]; the model was able to correctly scansion almost all the verses that we used as input. The model predictions were almost always correct, the only errors were related to the metrical type prediction function. This suggests that the model has correctly learned to predict the metrical pattern of Old Saxon verses. We tested the model also with some verses from *The Battle of Brunanburh* [21], but the results were quite different from the previous test. The model predicted almost all verses wrongly, suggesting that the few correctly predicted patterns were just random correct predictions, and the model is not able to generalize well over different distributions.



#### 4. FUTURE PERSPECTIVES

We presented a first attempt to create a tool for the automatic scansion of Old English and Old Saxon poetry. Despite the fact that the model is not able to predict the Old English meter, based on the results we got for Old Saxon we can conclude that our attempt was all in all successful. The dataset on which the model was trained, will be published on GitHub.

There remain many aspects of this project that can be implemented in future studies. Firstly, it could be useful to implement a specific syllabifier for both Old Saxon and Old English, which may lead to a more accurate syllabification of the verses. Secondly, future studies could focus also on the metrical type prediction function which, as shown by the results on the *Old Saxon Genesis*, sometimes led to incorrect associations with the metrical type. Thirdly, the incapacity of our model to predict the metrical pattern of Old English verses is probably due to the issue of domain shift in which the distribution of the test set is very different from the distribution of the new data passed to the model. Thus, a possible future implementation could be either to enrich the dataset with further verses from Old English poems, or to develop a new model trained with Old English verses.

#### REFERENCES

- [1] Agirrezabal Zabaletak, Manex. “Automatic Scansion of Poetry”. PhD diss., Universidad del Pais Vasco, 2017.
- [2] Agirrezabal, Manex and Astigarraga, Aitzol and Bertol, Arietta and Mans, Hulden. “ZeuScansion: A tool for scansion of English poetry”. 2016. *Journal of Language Modelling*. Vol. 4, No. 1. 3-28.
- [3] Beaudouin, Valérie and Yvon, Francois. “The Metrometer: A Tool for Analysing French Verse”. 1996. *Literary & Linguistic Computing*. Vol. 11, n. 1. 23-32. <http://dx.doi.org/10.1093/lc/11.1.23>
- [4] Bliss, Alan J. *The Meter of Beowulf*. Oxford: Basil Blackwell, 1967.
- [5] Gervás, Pablo. “A Logic Programming Application for the Analysis of Spanish Verse”. 2000. *Computational Logic (CL): Lecture Notes in Computer Science*. Vol. 1861. [https://doi.org/10.1007/3-540-44957-4\\_89](https://doi.org/10.1007/3-540-44957-4_89)
- [6] Gippert, Jost. 2003. *Genesis*. Based on Burkhard, Taeger and Verlag, Max Niemeyer. Tübingen. 1996. <https://titus.uni-frankfurt.de/texte/etcs/germ/asachs/genesis/genes.htm>
- [7] Griffo, Umberto. n.d. “Focal Loss”. *Categorical Focal loss*. <https://github.com/umbertogriffo/focal-loss-keras>
- [8] Hayes, Bruce and Wilson, Colin and Shisko, Anne. “Maxent grammars for the metrics of Shakespeare and Milton”. 2012. *Language*, 88. 691-731. <http://dx.doi.org/10.2307/23357550>
- [9] Hench, Christopher and Estes, Alex. 2015. “SonoriPy”. *SonoriPy*. 2015. <https://github.com/alexestes/SonoriPy>
- [10] Hochreiter, Sepp and Schmidhuber, Jürgen. “Long Short-Term Memory”. 1997. *Neural Computation*. 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [11] Keras Team. 2022. “Keras: Deep Learning for humans”. *Keras*. Last modified on November 14, 2022. <https://github.com/keras-team/keras>. Accessed February 17, 2023.
- [12] Keras. n.d. “Keras Tuner”. *Keras Tuner*. [https://keras.io/keras\\_tuner/](https://keras.io/keras_tuner/). Accessed February 17, 2023.
- [13] Logan, Harry M. “Computer Analysis of Sound and Meter in Poetry”. 1988. *College Literature* 15, no. 1. The Johns Hopkins University Press. 19-24.
- [14] Lühr, Rosemarie and Gippert, Jost. 1955. *Heliand*. Based on Taeger, Burkhard and Verlag Max Niemeyer. Tübingen. 1984. <https://titus.uni-frankfurt.de/texte/etcs/germ/asachs/heliand/helia.htm>
- [15] Pede Certo. *Musique Deoque*. 2022. “Pede certo”. *Pede certo*. Last modified on August 02, 2022. <https://www.pedecerto.eu/public/index>
- [16] SeatGeek. 2020. “FuzzyWuzzy”. *Process.py*. February 13, 2020. <https://github.com/seatgeek/fuzzywuzzy/blob/master/fuzzywuzzy/process.py>. Accessed February 23, 2023
- [17] Sievers, Edward. “Zur Rhythmik des germanischen Alliterationsverses. I.”. 1885. *10*. Vol. 2. 209-314.
- [18] Suzuki, Seiichi. *The Meter of Old Saxon Poetry. The Remarkings of Alliterative Tradition*. Cambridge: Brewer, 2004.
- [19] Text Encoding Initiative (TEI). n.d. “TEI Text Encoding Initiative”. *Text Encoding Initiative*. <https://tei-c.org/>. Accessed February 17, 2023.
- [20] Tueller, Michael A. “HYLAS: A new metrical search tool for Greek and Latin Poetry”. 2022. *Culture Digitali: Intersezioni, Filosofia, Arti, Media*. Università del Salento. 26-29.
- [21] University of Calgary. n.d. “The Battle of Brunanburh”. *The Battle of Brunanburh*. <http://people.ucalgary.ca/~mmcilli/ASPR/a10.1.html>

# Diachronic and diatopic word embeddings from British historical newspapers

Nilo Pedrazzini<sup>1</sup>, Barbara McGillivray<sup>2</sup>

<sup>1</sup> The Alan Turing Institute / University of Oxford, United Kingdom – [npedrazzini@turing.ac.uk](mailto:npedrazzini@turing.ac.uk)

<sup>2</sup> King’s College London, United Kingdom – [barbara.mcgillivray@kcl.ac.uk](mailto:barbara.mcgillivray@kcl.ac.uk)

## ABSTRACT

This poster presents a new resource for the study of diatopic semantic variation in historical texts consisting of word embedding models trained on historical British newspapers for five geographical regions. We discuss the embedding models and present an analysis on the lexicon of mechanisation in 19th-century English. As an application of our models, we show how different results were obtained from running a changepoint detection algorithm on the embeddings for the North and South of England, historically corresponding to a major socio-political split in Britain. This suggests how the semantic change of words related to the mechanisation process following the Industrial Revolution did not occur at the same pace across British regions. Our methods can be applied to other languages and historical texts, and our resources can be reused to investigate other questions related to semantic change in 19th-century English.

## KEYWORDS

Historical corpora; word embeddings; semantic variation; semantic change

## 1. BACKGROUND AND OBJECTIVES

Interpreting meaning expressed in text is a fundamental aspect of humanistic research. With the recent growth in the availability of historical texts in digital format, researchers now have the opportunity to mine these collections at scale. Computational methods allow us to conduct a variety of semantic analyses on large textual collections, including detecting evolving word meanings over time [20] and tracing conceptual change [4], which can support research on socio-cultural phenomena (e.g. [11]). State-of-the-art techniques rely on word embeddings to generate low-dimensional vector representations of words from words’ co-occurrence data [8, 16] which capture important semantic properties of words, including similarity and analogy relations. Recent years have witnessed a growth in the number of research projects generating diachronic word2vec embeddings [14] from historical texts. Researchers have released word embeddings trained on various diachronic corpora [5, 6, 7, 10]. Diatopic variation has been the object of much research in variational linguistics, with most studies focussed on synchronic data (e.g. [3, 18, 21]), but less attention has been devoted to its quantitative study in diachronic contexts. This poster presents a new resource for the study of diatopic semantic variation in historical texts consisting of word embedding models trained on historical British newspapers for five different geographical regions, building on the experiment in [15] by adding the diatopic dimension to the diachronic one. The objectives of this study are to discuss the embedding models and to outline the results of our analysis on the lexicon related to mechanisation in 19th-century English. Our methods are general enough to be relevant to research on other languages and historical texts.

## 2. DATA AND METHODS

We used a corpus of historical British newspapers comprising around 4.6 billion tokens and spanning the period between 1801 and 1920. The corpus includes titles specifically selected for the Living with Machines project<sup>1</sup> (2.3 billion tokens) and selected titles from the British Library’s Heritage Made Digital digitization project (further 2.3 billion tokens).<sup>2</sup> Using the place and year of publication of each newspaper, we divided the corpus into two subcorpora containing articles published in two broad geographical regions, North and South England, historically corresponding to one of the main socio-political divides in Britain, and we split each geographical subcorpus into 10-year slices.<sup>3</sup> Because of the size of the corpus, the texts underwent minimal pre-processing (lowercasing, punctuation and stopword removal) and no lemmatization. We then trained Word2Vec [14, 17] models for each decade in each geographical subcorpus, which we release as a resource

<sup>1</sup> <https://www.turing.ac.uk/research/research-projects/living-machines>

<sup>2</sup> <https://www.bl.uk/projects/heritage-made-digital>

<sup>3</sup> Training of diachronic models for additional British regions (Midlands, Scotland, and Wales) is also underway thanks to additional historical newspaper data from the British Newspaper Archive.

for the community.<sup>4</sup> We also use the embeddings to trace the semantics of lemmas related to mechanisation across different decades. We aligned the semantic spaces via Orthogonal Procrustes [19] and used the cosine similarity between vectors across different decades to measure their semantic shift and the pruned exact linear time (PELT) algorithm [9] to detect potential semantic changes in each geographic subcorpus. The results from the overlapping time-slices between the subcorpora<sup>5</sup> were compared against each other to assess whether their semantic shift occurred virtually simultaneously across the two regions or whether some degree of diatopic variation could be posited.

### 3. RESULTS

Word	North England	South England
bulbs	1860s	-
cars	1860s	-
machines	-	1860s
match	1860s	1840s
matches	1860s	-
stamp	1860s	1860s
stamps	-	1840s
stock	-	1860s
trade	1860s	-
trolley	<i>not in vocabulary</i>	1850s

Table 1. Changepoints detected in the North-England and South-England subcorpora.

Our preliminary results show differences between the changepoints detected by the PELT algorithm for the North and South of England. While in some cases, like *match* and *stamp*, a changepoint was detected for both regions, for other words, such as *machines* and *stock*, a changepoint was only detected for the South, whereas for others, like *trade*, *bulbs* and *cars*, only for the North. For some words, a changepoint was detected in a region either only in the singular or the plural form, as in *match* (but not *matches*) in the South or *stamp* (but not *stamps*) in the North of England; in other cases a changepoint was detected earlier for one of the two forms. These subtle differences may have to do with the different usages and therefore different triggers for semantic change associated with some words in the singular and plural, for example when they are used as plural generics [12, 13], as opposed to referring to multiple instances of a concrete object. Moreover, even for words with a potential changepoint in both regions, the decade in which the shift occurred may differ: for *match*, for instance, a changepoint was detected later in the century for the North than for the South. Existing corpus query tools such as BNClab [1] enable lexical analyses of language usage across time; however, they do not offer a semantic search functionality, therefore they do not allow us to look for evidence of such semantic shifts in diatopic variation. To interpret these results and evaluate whether they correspond to historically driven intuitions, we plan to extract the nearest neighbours of the words for which a changepoint was detected in any of the models in the decades before and after the changepoints. The words found among the nearest neighbours should help us identify the type of semantic change which occurred for a given word. The shift in nearest neighbours (to be expected given the detected changepoints) can also be compared across different regions, to check whether the same kind of shift occurred across the board or whether, besides the difference in changepoint, a further difference in the type of semantic change can be observed.

<sup>4</sup> The models can be found at <https://doi.org/10.5281/zenodo.7892460>; the associated code to train diachronic word embeddings can be found at <https://github.com/Living-with-machines/DiachronicEmb-BigHistData>

<sup>5</sup> The corpus used for this preliminary experiment is imbalanced in temporal coverage: articles from North England cover the span 1830s-1910s, whereas those from South England 1800s-1880s. Additional models covering the remaining decades for each region are also underway thanks to the British Newspaper Corpus. For the purpose of this poster, we only present results from the portion of the century for which both North and South England have some temporal coverage (i.e. 1830s-1880s).

## 4. ACKNOWLEDGEMENTS

Work for this paper was carried out as part of *Living with Machines*. This project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC grant AH/S01179X/1), with The Alan Turing Institute, the British Library and King's College London, East Anglia, Exeter, and Queen Mary University of London. Newspaper digitisation by the British Library, undertaken as part of the Living with Machines project, was funded by the AHRC.

## REFERENCES

- [1] Brezina, V., Gablasova, D. & Reichelt, S. (2018). "BNClab". <http://corpora.lancs.ac.uk/bnclab> [electronic resource, last accessed May 2023], Lancaster University.
- [2] Camacho-Collados, Jose and Mohammad Taher Pilehvar. "From word to sense embeddings: a survey on vector representations of meaning," *Journal of Artificial Intelligence Research* 63, no. 1 (2018): 743–788.
- [3] De Pascale, Stefano. Token-based vector space models as semantic control in lexical lectometry. KU Leuven: PhD thesis, 2019.
- [4] Fokkens, Antske, Serge ter Braake, Isa Maks and Davide Ceolin. "On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change," *Proceedings of Drift-a-LOD@EKAW* (2016).
- [5] Grayson, Siobhán, Maria Mulvany, Karen Wade, Gerardine Meaney and Derek Greene. "Novel2Vec: Characterising 19th Century Fiction via Word Embeddings," *Proceedings of the 24th Irish Conference on Artificial Intelligence and Cognitive Science*, (2016): 20–21.
- [6] Hamilton, William L., Jure Leskovec and Dan Jurafsky. (2016). "HistWords: Word Embeddings for Historical Text". <https://nlp.stanford.edu/projects/histwords/> [electronic resource, last accessed May 2023].
- [7] Hosseini, Kasra, Kaspar Beelen, Giovanni Colavizza and Mariona Coll Ardanuy. "Neural Language Models for Nineteenth-Century English," *Journal of Open Humanities Data* 7, no. 22 (2021). <http://doi.org/10.5334/johd.48>
- [8] Joulin, Armand, Edouard Grave and Piotr Bojanowski Tomas Mikolov. "Bag of Tricks for Efficient Text Classification," *Proceedings of EACL 2017*, (2017): 427–431.
- [9] Killick, Roberta, Paul Fearnhead and Idris A. Eckley. "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association* 107, no. 500 (2012): 1590–1598.
- [10] Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde and Slav Petrov. "Temporal Analysis of Language through Neural Language Models," *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, (2014): 61–65.
- [11] Kozłowski, Austin C., Matt Taddy and James Evans. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," *American Sociological Review* 84, no. 5, (2019): 905–949. <https://doi.org/10.1177/0003122419877135>
- [12] Leslie, Sarah J., Sangeet Khemlan, Sandeep Prasada and Sam Glucksberg. "Conceptual and linguistic distinctions between singular and plural generics," *Proceedings of the 31st Annual Cognitive Science Society*, (2009): 479–484.
- [13] Mari, Alda, Claire Beyssade and Fabio Del Prete. *Genericity*. Oxford: Oxford University Press, 2012.
- [14] Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. "Efficient estimation of word representations in vector space," *Proceedings of Workshop at the International Conference on Learning Representations*, (2013).
- [15] Pedrazzini, Nilo and Barbara McGillivray. "Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers," *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, (2022): 85–95.
- [16] Pennington Jeffrey, Socher Richard, Manning Christopher D. "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2014): 1532–43.
- [17] Řehůřek, Radim and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (2010): 45–50.
- [18] Ruetten, Tom, Dirk Speelman and Dirk Geeraerts. "Measuring the lexical distance between registers in national varieties of Dutch," *Proceedings of the International Conference on Pluricentric Languages*, (2011): 541–554.
- [19] Schönemann, Peter H. "A generalized solution of the orthogonal procrustes problem," *Psychometrika* 31, (1966): 1–10.
- [20] Tahmasebi, Nina, Lars Borin, Adam Jatowt, Yang Xu and Simon Hengchen. (Eds.). *Computational approaches to semantic change* (Vol. 6). Language Science Press, 2021.
- [21] Wieling, Martijn, Simonetta Montemagni, John Nerbonne and Harald R. Baayen. "Lexical Differences between Tuscan Dialects and Standard Italian: Accounting for Geographic and Sociodemographic Variation Using Generalized Additive Mixed Modeling," *Language* 90, no. 3 (2014): 669–92.

# GPT-3 vs. Delta. Applying stylometry to large language models

Simone Rebora<sup>1</sup>

<sup>1</sup> Johannes Gutenberg University Mainz, Germany – simone.rebora@uni-mainz.de

## ABSTRACT

This paper tests the ability of large language models to deceive stylometric approaches in authorship attribution. A corpus of ten English authors is used as a reference point, while GPT-3 is asked to generate texts that imitate their style. After having defined a baseline for the efficiency of stylometric methods on human-generated texts, a series of analysis is performed on the artificially generated texts. Results show the inability of GPT-3 to deceive stylometry and allow a quantitative analysis of its distinctive linguistic features. Preliminary results are also presented for ChatGPT, indicating the efficiency of stylometry in detecting its authorial fingerprint.

## KEYWORDS

Stylometry; Authorship attribution; Large language models; GPT-3; ChatGPT

## 1. INTRODUCTION

The growing interest towards large language models (LLMs) such as [GPT-3](#), [ChatGPT](#), and [the new Microsoft Bing](#) (March 2023), has been accompanied by the preoccupation if the text generated by these systems can be identified automatically, so to control its possible misuse. Works like the one by [5] have already aimed at identifying the features that allow such a recognition, while online services such as [GPT-Zero](#) offer the possibility to easily test its efficiency. Even if preliminary results are promising, [12] warns about the necessity of moving beyond traditional stylometric methods to identify artificially generated text. Apart from limited applications like [8], still, no extensive study has been dedicated to the automated recognition of LLMs in literary studies.

This paper tries to fill this gap by testing to what extent one language model (GPT-3) is able to deceive stylometric approaches in authorship attribution. In fact, while stylometry has clearly demonstrated its potential in recognizing the authorial fingerprint of literary authors [1], examples have already been shown where authors were able to deceive the algorithm, through the practices of pastiche and imitation (see for example Joyce in [9]).

## 2. DEFINING THE BASELINE

In order to perform such an analysis, a baseline needs to be defined, indicating the efficiency of stylometric methods in the attribution of text written by humans.

The [ELTeC corpus of English literature](#) (created in the context of the *Distant Reading for European Literary History* COST Action) was selected for this aim, because of its balanced structure and detailed documentation [11]. 30 texts were extracted from it, after having identified the authors that contributed more than two texts to the collection. Out of these, 20 texts (two per author) constituted the training set, while 10 (one per author) constituted the test set.

A series of preliminary analyses<sup>1</sup> was then performed on the training set to identify the best performing stylometric features. Based on the research by [4], the Cosine Delta distance measure (also known as Würzburg Delta) was kept constant, while most frequent words (MFWs) were varied. As Figure 1 shows, the dendrogram reaches stability at 150 MFWs, clearly distinguishing the different authors in separated clusters. While stability is also confirmed by higher MFW selections, the value of 150 was selected because it limits the analysis mainly to function words, already identified as the possible groundwork for the success of stylometry in authorship attribution [7].

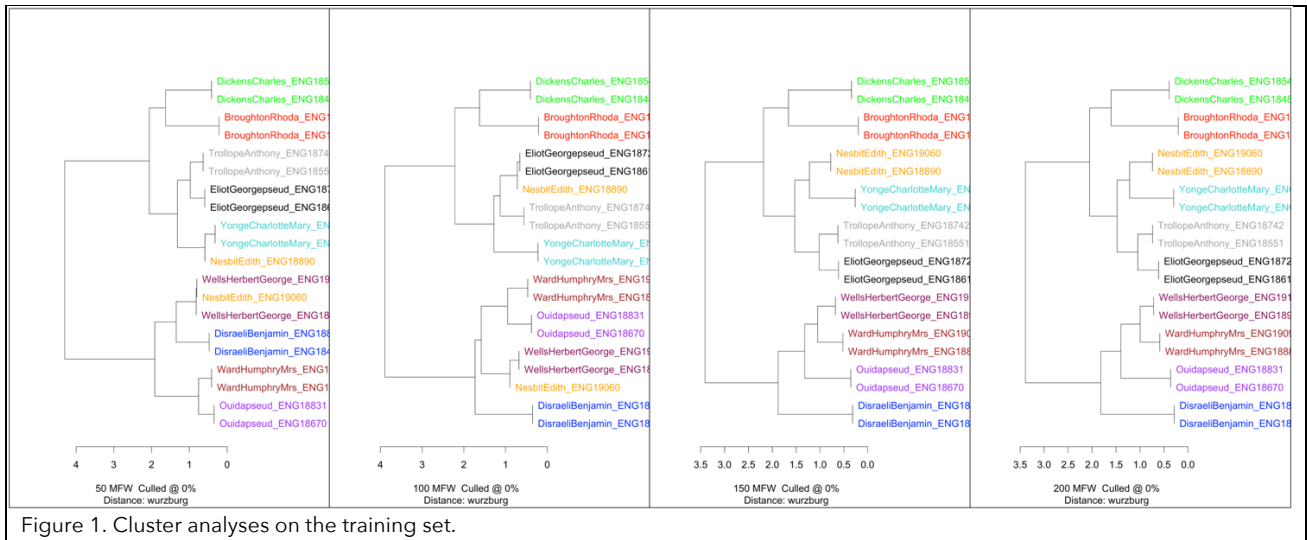
After having identified the stylometric features, the baseline was defined by comparing the training set with the test set. The 10 texts in the test set were split into chunks of 5,000 words (identified as the minimum length for a reliable stylometric analysis [2]), thus producing a total of 345 text samples. These samples were compared one by one with the training set, verifying if the stylometric analysis were able to cluster them close to the right author. Overall, 83.5% of the analyses were

---

<sup>1</sup> All analyses were performed using the *stylo* R package [3]. Scripts and datasets are available on GitHub: [https://github.com/SimoneRebora/GPT3\\_stylometry](https://github.com/SimoneRebora/GPT3_stylometry)



successful, with a success rate always higher than 68.7% for single authors. These values were therefore taken as a baseline for the following experimentation with GPT-3.



### 3. CREATING AN ARTIFICIAL TEST SET WITH GPT-3

To instruct GPT-3 into imitating the style of the 10 authors in the training set, a series of strategies were implemented via the [OpenAI API](#). An overview is provided by Table 1, which shows the four features that were varied and combined:

- the name of the author to be imitated;
- a prompt giving instructions on how to generate the new text;
- the model to be used (being DaVinci the most advanced and Ada the most basic);
- the temperature of the model (intended as how much the model diverges from a deterministic behavior, thus becoming more and more “creative”).

All combinations between these four features were tested, thus generating a total of 320 different configurations.

author_name	prompt	model	temperature
Rhoda Broughton	Write a chapter of a novel in the style of <author_name>	text-davinci-003	0.1
Charles Dickens			
Benjamin Disraeli			
George Eliot	Write a novel with a complex structure. It should include many events, dialogues, and descriptions. Write like <author_name>	text-curie-001	0.9
Edith Nesbit			
Ouida			
Anthony Trollope	Write a novel by imitating <author_name>	text-babbage-001	0.9
Mrs. Humphry Ward			
Herbert George Wells			
Charlotte Mary Yonge			
	Write a story as if it were written by <author_name>. Give it the structure typical of <author_name>'s narratives	text-ada-001	

Table 1. GPT-3 text generation strategy overview.

One issue to be dealt with was that of text length. Independently from the prompt, in fact, the models always tended to generate short texts, with an average length of 330 tokens. The inability of GPT-3 to generate long texts (or even to understand what text length means) [has already been discussed](#): indeed, adding to the prompts indications such as “write a long text” or “write 1,000 words” did not influence the results. For this reason, it was decided to just not include them, giving instead preference to structural indications (such as: “include many events, dialogues, and descriptions”). To reach the minimum of 5,000 words requested for a reliable stylometric analysis, text generation was repeated multiple times for each configuration, until the total amount of generated text surpassed the limit of 6,000 tokens (roughly corresponding to 5,000 words). All the generated texts were then concatenated into single text samples. While such a practice has already been suggested by [2] as valid in stylometric analysis, a simple procedure was devised to confirm its validity. First, each text in the original test set was split into 250-word chunks (roughly corresponding to 330 tokens) and artificially

recomposed by randomly selecting and joining 20 of them. The resulting textoids, being 5,000-word long, were then analyzed stylometrically by applying the same procedure for the definition of the baseline as described in paragraph 2. The analysis was repeated 30 times for each author, thus resulting in a total of 300 different analyses. With this new configuration, success rate increased even to 97.3%, finally showing how such a procedure for text generation might even strengthen the authorial signal.

A preliminary qualitative analysis of the artificially generated corpus confirmed the impression that GPT-3 correctly interpreted the task. In the case of Charles Dickens, for example, texts were inhabited by the most famous characters of his novels, even with evident cases of imitation (like almost verbatim repetitions of the incipit of *A Tale of Two Cities*).

#### 4. STYLOMETRIC ANALYSIS

The analysis of the artificially generated corpus was performed by once again adapting the procedure described in paragraph 2: each text sample was compared to the training set by using Cosine Delta distance and 150 MFW, verifying if it clustered together with the author it tried to imitate.

Overall, only 16.8% of the samples were attributed to the imitated author, thus indicating the substantial inability of GPT-3 to deceive the stylometric analysis. Some features showed higher efficiency than others (22.5% for the “Write a chapter of a novel in the style of <author\_name>” prompt; 25% for the DaVinci model; 19.4% for the 0.9 temperature), but still much below the baseline.

Figure 2 shows the results for different imitated authors, indicating how the partial success for some of them might depend more on the intrinsic characteristics of GPT-3’s “writing style” (more similar to the one of these authors—in particular Herbert George Wells), than on its efforts to imitate their style.

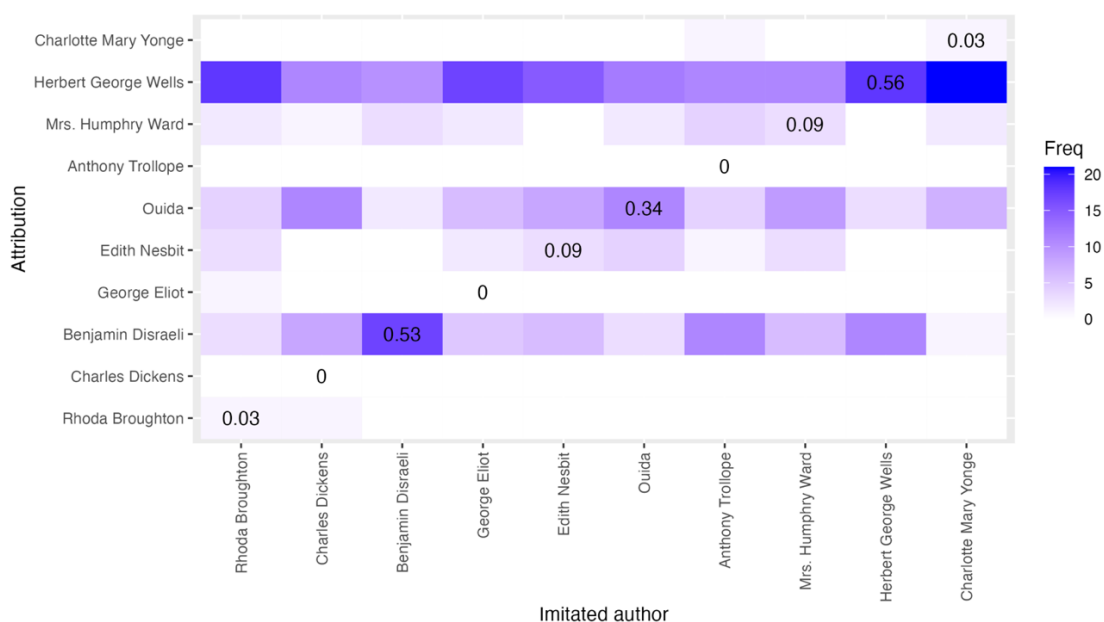


Figure 2. Attributions per imitated author (scores on the diagonal indicate the efficiency per author).

This result advises for a more detailed analysis of GPT-3’s style. As the dendrogram in Figure 3a shows, in fact, all the texts created by GPT-3 tend to group in a separated cluster, independently from the imitated author. The effect of imitation emerges only marginally when reducing the analysis to the artificially generated texts: in this case, the strongest stylometric signal seems to be generated by the different models, with DaVinci distinguishing itself more clearly (see Figure 3b).



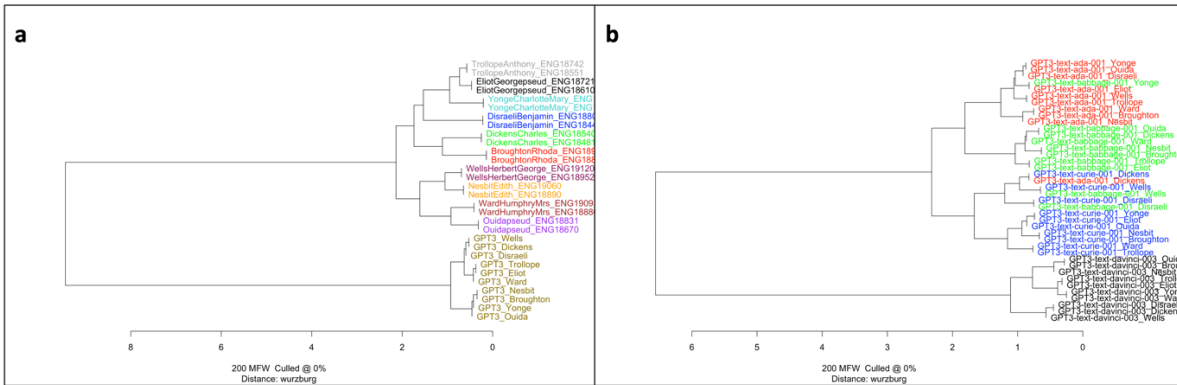


Figure 3. Cluster analysis of GPT-3 texts.

To dig deeper into the distinctive characteristics of GPT-3’s language, a procedure was developed by working on the Zeta scores, which constitute the groundwork for approaches like Cosine Delta. As shown by [10], in fact, the Delta method automatically looks for patterns in the Zeta scores (indicating how much the frequency of a word in a text deviates from its average frequency in a corpus). One possibility for (partially) reverse engineering the Delta procedure is that of looking for the words in a group of clustered texts that show the lowest variance in the Zeta scores (thus indicating an internally coherent behaviour, which is at the same time distinct from the rest of the corpus).

Figure 4 shows the results of such a procedure when applied (a) to all the GPT-3 texts versus the training set and (b) to the GPT-3 DaVinci texts versus the training set. A few aspects need to be noticed here. First, GPT-3 voices distinguish themselves only by the underuse (and not the overuse) of words. It therefore seems that GPT-3 applies a form of self-restraint in text generation. In particular, DaVinci distinguishes itself through the underuse of negation (“don”, “not”) and of modal verbs such as “might” (generally underused by all GPT-3 models) and “should”. While further research is indeed required, this phenomenon seems to be caused by a tendency of GPT-3 towards a more assertive tone, which aims at synthesis and avoids complex constructions.

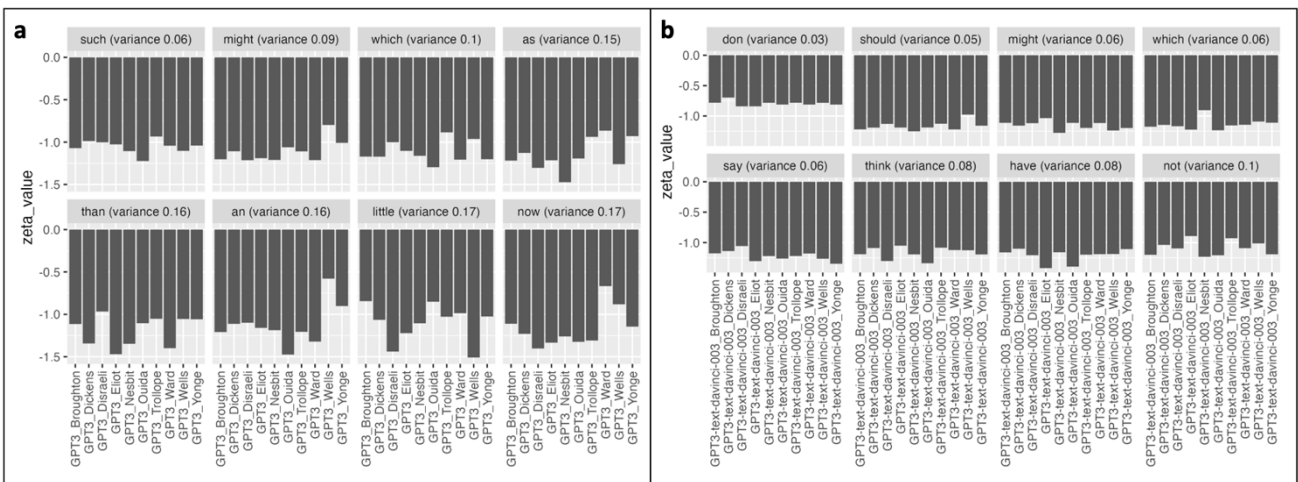


Figure 4. Most distinctive GPT-3 words.

See for example this passage, generated by the DaVinci model (temperature 0.1) when answering to the prompt “Write a chapter of a novel in the style of George Eliot”:

John was walking along the path, his feet crunching on the gravel beneath him. He had been walking for hours, and he was tired, but he was determined to keep going. He had been walking for days, ever since he had left his home in search of something better. [...] He kept walking, and eventually, he came to a small village. He stopped and looked around, and he could see that the village was poor, but it was also full of life. He could see people walking around, talking and laughing. He could see children playing in the streets. He felt a sense of hope, and he knew that this was the place he had been looking for. He had found something better.

Note the insisted repetition of “he + verb” construct and the use of short, simple sentences which bring the narrative tension to a rather hasty resolution. This passage exemplifies a tendency which is dominant in most of the artificially-generated

texts, able to identify at least some of the defining elements in one author’s literary production (on the quest for “something better” in George Eliot, see [6]), but expressing it in a dry and repetitive manner, which could satisfy a superficial reading, but not any deep reading or critical interpretation.

### 5. PRELIMINARY RESULTS WITH CHATGPT

GPT-3 was chosen for this study because, at the time of writing, it provided an API which allowed extensive testing. While a similar analysis is not (yet) possible with the more powerful ChatGPT, a preliminary experiment was attempted to verify if the above results could be confirmed.

Two strategies were applied with only one author. In the first attempt, the initial prompt: “Write the incipit of a novel in the style of Charles Dickens”; was followed by the repeated prompt: “Please go on in the style of Charles Dickens”. This approach was only marginally unsuccessful, because ChatGPT concluded the story already at the third interaction, continuously reshaping the ending in the following ones. For this reason, a second, more complex approach was adopted. A first prompt: “Set up a plan for writing the chapter of a novel in the style of Charles Dickens. It should contain actions, descriptions, and dialogues, in the proportion typical of Dicken’s narratives. It should be divided into 10 [or 15] parts”; was followed by the request to write the parts one by one. This approach was more successful, as it produced three coherent stories of a length between 1,300 and 3,300 words. In any case, more than 5,000 words were collected for both attempts. Figure 5 confirms the inability of ChatGPT to deceive the stylometric approaches and it also indicates the possible distinctiveness of its voice, placing itself close to Ouida.

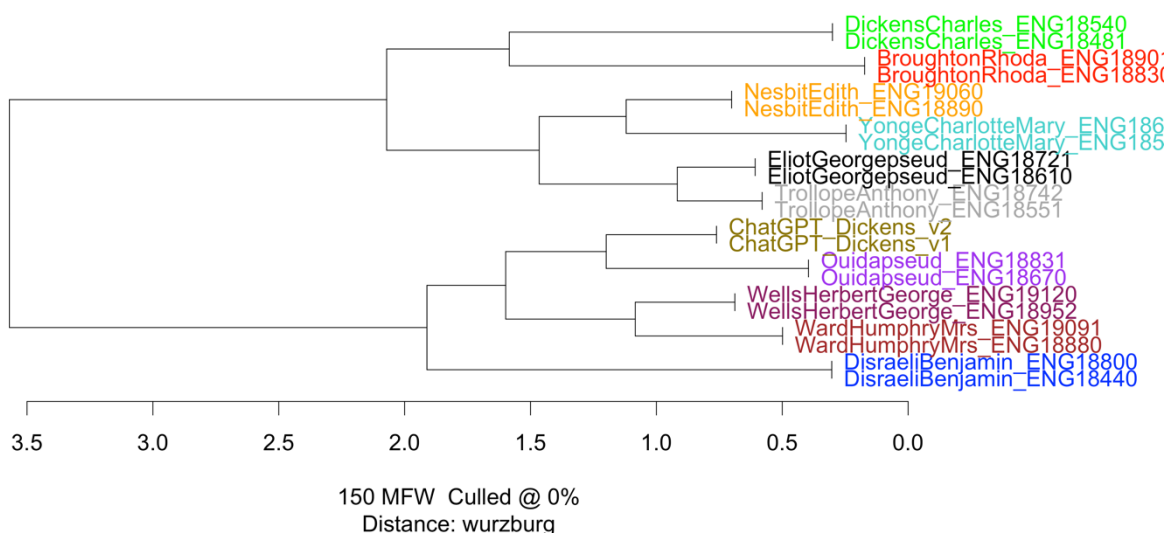


Figure 5. Cluster analysis of ChatGPT texts.

### 6. CONCLUSIONS

One of the main issues in carrying out such a research is that LLM technology evolves at an increasing speed. What is not yet available at the time of writing (e.g., an API for ChatGPT, or the much advertised [GPT-4](#)) might be published in the very near future. For this reason, the results presented here should be considered as just a screenshot of a continuously evolving landscape. Still, the coherence between the results obtained with GPT-3 and with ChatGPT suggests how LLMs are still far from effectively imitating the style of literary authors.

A final note should be then added on the many alternative experimental designs that could have been applied for this study. First, different features (such as character n-grams or parts of speech) and approaches (such as machine learning) could have been used for the stylometric analysis: this study focused just on MFWs and distance measures as they allowed a higher explainability of the results, but further research with different approaches is indeed needed to test the full potential of stylometry. Another possible different design could have involved contemporary literary authors. Indeed, one might argue that LLMs have more difficulty in imitating the style of authors from the past, being mainly trained on contemporary text. While this is a sharable concern, copyright restrictions make such an experimentation much less feasible. The free availability of corpora like ELTeC is what made this study possible in the first place. Another option would have been that of fine-tuning the models on the texts by the imitated authors: while this should probably improve the efficiency of the

models, it might tell us less about their intrinsic characteristics. For this reason, this study avoided any fine-tuning, just testing the models in their “raw” potential. Finally, the “art of prompting” has recently affirmed itself as one of the fundamental skills in the use of LLMs. Here, one possible weakness of this study could be the effectiveness of the textual prompts, which were kept rather simple. While simplicity was chosen to avoid instructing too much the models and to test their zero-shot efficiency, further studies should explore how more sophisticated prompts could generate better imitations of authorial style.

Notwithstanding all these issues and limitations, the results presented here could still be used as the groundwork for further research and experimentation, profiting of the confrontation between stylometry and LLMs to better understand them both.

#### REFERENCES

- [1] Burrows, John. “‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship.” *Literary and Linguistic Computing* 17, no. 3 (2002): 267–87. <https://doi.org/10.1093/lc/17.3.267>
- [2] Eder, Maciej. “Does Size Matter? Authorship Attribution, Small Samples, Big Problem.” *Digital Scholarship in the Humanities* 30, no. 2 (2013): 167–82. <https://doi.org/10.1093/lc/fqt066>
- [3] Eder, Maciej, Jan Rybicki, and Mike Kestemont. “Stylometry with R: A Package for Computational Text Analysis.” *The R Journal* 8, no. 1 (2016): 107–21.
- [4] Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. “Understanding and Explaining Delta Measures for Authorship Attribution.” *Digital Scholarship in the Humanities* 32, no. suppl\_2 (2017): ii4–16. <https://doi.org/10.1093/lc/fqx023>
- [5] Fröhling, Leon, and Arkaitz Zubiaga. “Feature-Based Detection of Automated Language Models: Tackling GPT-2, GPT-3 and Grover.” *PeerJ Computer Science* 7 (2021): e443. <https://doi.org/10.7717/peerj-cs.443>
- [6] Hornback, Bert G. “The Moral Imagination of George Eliot.” *Papers on Language and Literature* 8, no. 4 (1972): 380–94.
- [7] Kestemont, Mike. “Function Words in Authorship Attribution. From Black Magic to Theory?” In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, 59–66. Gothenburg, Sweden: Association for Computational Linguistics, 2014. <http://aclweb.org/anthology/W/W14/W14-0908.pdf>
- [8] Lawson, Rebecca. “GPT-2: Girl Detective Analyzing AI-Generated Nancy Drew with Stylometry.” *IPHS 200: Programming Humanity*, October 1, 2020. [https://digital.kenyon.edu/dh\\_iphs\\_prog/32](https://digital.kenyon.edu/dh_iphs_prog/32)
- [9] Reborá, Simone. “Encyclopedic Novel Revisited. Joyce’s Role in a Disputed Literary Genre.” *Joyce Studies in Italy* 19 (2017): 147–68.
- [10] “Stylometry and Reader Response. An Experiment with Harry Potter Fanfiction.” In *AIUCD 2022 - Proceedings*, edited by Fabio Ciraci, Giulia Miglietta, and Carola Gatto, 30–34. Bologna: AIUCD, 2022. <http://amsacta.unibo.it/6848/>
- [11] Schöch, Christof, Roxana Patras, Tomaz Erjavec, and Diana Santos. “Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives.” *Modern Languages Open* 1 (2021): 25. <https://doi.org/10.3828/mlo.v0i0.364>
- [12] Schuster, Tal, Roi Schuster, Darsh J. Shah, and Regina Barzilay. “The Limitations of Stylometry for Detecting Machine-Generated Fake News.” *Computational Linguistics* 46, no. 2 (2020): 499–510. [https://doi.org/10.1162/coli\\_a\\_00380](https://doi.org/10.1162/coli_a_00380)

# EcoLet. Ecotemario della Letteratura italiana

Daniele Silvi<sup>1</sup>

<sup>1</sup> Università degli Studi di Roma Tor Vergata – silvi@lettere.uniroma2.it

## ABSTRACT

Questo intervento vuole portare all'uditorio una serie di riflessioni sull'organizzazione della conoscenza e sull'impiego di modelli computazionali per la critica letteraria e tematica nello specifico. Alla luce della fortunata esperienza di progetti che hanno proposto una annotazione tematica come *Memorata Poetis. Ricorrenze lessicali e tematiche nella versificazione epigrafica e letteraria*, avvalendosi di uno specifico temario che ha, a sua volta, richiesto un grande lavoro di progettazione e di sviluppo, si intende qui proporre la progettazione e lo sviluppo di un temario che abbia le stesse finalità ma che sia centrato sui temi della ecologia letteraria.

## PAROLE CHIAVE

Ecocritica; ontologia; Bert; temario; ecologia

## 1. INTRODUZIONE

L'obiettivo centrale di questa proposta si fonda su un piano strettamente teorico letterario e metodologico. Si cercherà infatti di procedere a una modellizzazione formale dei concetti di motivo, tema, topos e dei concetti connessi, nonché delle modalità con cui essi circolano nei macrotesti letterari di diversi periodi culturali, con esclusivo riferimento al rapporto tra Letteratura ed Ecologia [1]. L'idea è quella di definire concettualmente l'ambito della Eco-tematologia mediante l'uso di ontologie formali. A nostro avviso le tecnologie ontologiche del Web Semantico forniscono un apparato strumentale idoneo alla creazione di un repertorio tematico che non sia una pura enumerazione – come nei pur validi repertori a stampa – ma che al contrario si organizzi per multiple stratificazioni tipologiche e al contempo permetta una ricca rete di relazioni orizzontali tra classi e tra istanze di temi e motivi.

## 2. QUADRO TEORICO DI RIFERIMENTO

Nel corso degli ultimi due decenni la riflessione teorica sul letterario ha investito il rapporto uomo/ambiente e la sua mediazione tramite la scrittura. L'ecocritica, le cui origini si rintracciano nei lavori seminali di ambito nordamericano di W. Rueckert [2] e J. Meeker [3], è oggi una delle correnti innovative più interessanti e attuali del discorso culturale, nel senso etimologico del termine: in gradi cioè di agire performativamente sulla realtà sociale, in una fase in cui la transizione ecologica occupa l'agenda del discorso pubblico. Infatti, tale transizione, non priva di contraddizioni, potrà essere agita solo se la mediazione culturale sarà in grado di renderla senso comune.

Ma ridefinire il rapporto con la natura ha una portata ben più vasta che la semplice adozione di politiche di green economy. La crisi pandemica dell'ultimo biennio ha mostrato come la transizione ecologica sia ormai anche una necessità sociale che investe diversi aspetti dei comportamenti sociali. La maggiore sensibilità sul ruolo dell'ambiente di vita sulla salute dell'individuo inizia ad alimentare una nuova mentalità, una riconsiderazione della vita metropolitana e un bisogno di ricollocarsi nella dimensione del territorio, ovviamente propiziata dalle possibilità relazionali offerte dall'esistenza online.

In questo quadro ci proponiamo di indagare come la sfera della narrazione letteraria (intesa in una accezione vasta, che include la letteratura di genere, il cinema e la fiction televisiva) [4] dalla modernità industriale alla contemporaneità, abbia mediato la rappresentazione del territorio, del rapporto uomo/natura, dei processi di urbanizzazione e metropolizzazione, anche sulla spinta di fenomeni epocali di alterazione dell'ambiente naturale, tipici prima della Rivoluzione industriale e poi della globalizzazione. Il quadro teorico e storico sarà poi la base per definire pratiche di annotazione testuale su più vasta scala, e azioni con finalità culturali formative.

## 3. APPROCCIO METODOLOGICO

Su un campione di 100 testi narrativi della letteratura italiana dal 1830 al 1920, di 35 autori diversi, proveniente dalla collezione ELTeC, preventivamente codificati (dal sottoscritto, secondo il formato XML e lo schema Text Encoding Initiative - TEI per segmentare il testo in unità sub-testuali), è stata dapprima condotta una estrazione automatica dei temi, mediante l'applicazione al corpus dell'algoritmo di topic modelling più diffuso LDA [5]. Questo ha permesso di ottenere una prima lista – orientativa – su cui poi lavorare manualmente per un affinamento ed un arricchimento.

L'annotazione tematica in ambito letterario [6] infatti è comunemente considerata una operazione critica che, anche se supportata da tecnologie informatiche e computazionali [7], va essenzialmente eseguita in modo "manuale" poiché richiede il sostanziale intervento della capacità critica ed ermeneutica, nonché della competenza storico-letteraria, del ricercatore che la esegue. Recentemente tuttavia, nell'ambito del cosiddetto paradigma del distant reading [8] si sono ampiamente diffuse alcune tecniche di analisi statistico/probabilistica che cercano di operazionalizzare [9] la fase di individuazione dei cluster tematici presenti in un corpus testuale. In particolare ci riferiamo alle tecniche di topic modeling, ovvero l'estrazione dei cluster lessicali che caratterizzano un insieme di testi, e l'analisi delle loro distribuzioni [10] [11]. L'algoritmo utilizzato è quello di David Mimno, con ripetizioni incrementali su 100 topic (ulteriori e più dettagliati parametri saranno discussi in sede di presentazione).

Il temario che ne risulta deve quindi intendersi come uno strumento pensato per ulteriori lavori, sia di annotazione che di ricerca. Il primo passo è stato definire una tassonomia in cui i temi che man mano emergevano potessero trovare un posto anche in termini gerarchici. Questo mi ha portato ad individuare un primo livello di categorie tematiche. Le seguenti:

1. Relazione uomo-natura: in che modo gli esseri umani interagiscono con l'ambiente naturale,
2. Conservazione della natura: azioni e descrizioni orientate alla protezione dell'ambiente naturale, comprese le questioni legate alla conservazione della fauna selvatica, alla protezione dei parchi naturali e alla sostenibilità delle risorse naturali,
3. Biodiversità: descrizioni che facciano emergere le diversità degli ecosistemi, delle specie vegetali e animali,
4. Il cambiamento climatico: questo tema si concentra sui cambiamenti climatici e sui loro effetti sulla natura,
5. L'inquinamento ambientale: questo tema esplora le conseguenze dell'inquinamento sull'ambiente naturale,
6. Il rapporto tra cultura e natura: questo tema si concentra sulla relazione tra l'arte e la cultura e l'ambiente naturale,
7. La filosofia ecologica: questo tema esplora le teorie filosofiche che sostengono la protezione dell'ambiente naturale e la sostenibilità ambientale,
8. La giustizia ambientale: questo tema si concentra sui problemi di giustizia sociale e ambientale.

Una serie di livello inferiore di temi, che discendono da questi, sta venendo altresì individuata. Faccio qui solo alcuni esempi: tema di secondo livello di cui al punto 1 è la distruzione ambientale; del punto 2 è la preservazione della varietà della vita sulla Terra; del punto 3 sono tutte le diverse categorie di specie vegetali e animali che vengono (o potrebbero venire) descritte; del punto 4 sono la temperatura globale, l'innalzamento del livello del mare e la perdita di habitat naturali; del punto 5 sono gli effetti sulla salute umana e sulla qualità dell'aria e dell'acqua; del punto 6 (molto vasto) fanno parte la rappresentazione della natura nella letteratura, nelle arti visive e nella cultura popolare; del punto 7 (abbastanza complesso) fanno parte gli approcci ecocentrici e antropocentrici; del punto 8 fanno parte l'equità nella distribuzione delle risorse naturali e la responsabilità sociale delle imprese.

Per la ricerca tramite algoritmo LDA ho invece etichettato più genericamente una serie di categorie semantiche relative a ciascun tema, un insieme cioè di sostantivi che definiscano solo lessicalmente il tema in esame e che riportino appunto a categorie generali. Anche in questo caso cito solo alcuni esempi: 1) urbanizzazione, industrializzazione, alienazione, resilienza, simbiosi; 2) parchi nazionali, riserve naturali, biodiversità, ecosistemi, habitat; 3) specie, ecosistemi, habitat, endemismo, conservazione; 4) riscaldamento globale, effetto serra, cambiamenti nel clima, innalzamento del livello del mare, desertificazione; 5) inquinamento dell'aria, dell'acqua, del suolo, rifiuti tossici, smog; 6) romanticismo, pastoralismo, ambientalismo, antropocentrismo, ecocentrismo; 7) ecologia profonda, ecologia sociale, sostenibilità, bioetica, ecotopia; 8) disuguaglianza, discriminazione, sostenibilità, partecipazione, diritti ambientali, responsabilità, accessibilità. Naturalmente questo è un elenco di etichette che individuano temi che – nelle varie epoche e culture – sono stati denotati e connotati in maniera diversa. In altre parole i temi dell'ecologia sociale o dell'inquinamento del suolo, per fare degli esempi, possono assumere diverse connotazioni linguistiche a seconda del periodo e dell'orizzonte tecnologico di riferimento. La maggior parte del lavoro di descrizione del temario sarà quindi, nel futuro del progetto, quello di riunire tutte le marche tematiche che riconducano alla specifica etichetta differenziandole e descrivendole nel contesto storico-sociale che gli pertiene.

Come si può facilmente notare anche solamente da questi piccoli esempi, molti riferimenti lessicali sono intrecciati tra di loro e possono riferirsi a più temi. Inoltre molti termini (e il loro uso) sono subordinati al periodo storico ed – in parte – al genere letterario. La disambiguazione e l'ulteriore classificazione di tali termini nel contesto dei temi di appartenenza, rappresenta la parte del lavoro che è attualmente in fase di sviluppo.

Va di nuovo sottolineato in questa sede che per rintracciare una marca semantica ecologica in testi pre-novecenteschi e primo-novecenteschi, bisogna necessariamente tenere conto del panorama urbano, industriale e tecnologico di riferimento.

In tale contesto anche i termini paesaggistici della campagna assumeranno valenza ecologica. Lo stesso vale per la sfera dell'alimentazione e delle abitudini ad essa connesse (scarti alimentari, utilizzo di combustibili, ecc.). Sono quindi stati questi i primi termini che ho preso in considerazione e che hanno poi generato una serie di collegamenti. Per l'estrazione di questi termini 'primigeni' ho proceduto con una doppia logica: il rapporto tra la scelta manuale e i candidati proposti dagli algoritmi probabilistici è infatti di propedeuticità diretta. Dapprima ho studiato sia le concordanze sui testi che l'analisi tematica attraverso algoritmi di LDA, estraendo in contemporanea, manualmente da una lista di frequenze, i termini più direttamente collegati al tema ecologico e suddividendoli inoltre in categorie (secondo lo schema qui già esposto). La riflessione tecnica a questo punto ha posto la mia ricerca davanti ad un bivio metodologico: 1) proseguire per 'parole chiave' da utilizzare nella scansione degli altri testi, oppure 2) addestrare un modello linguistico (Bert, ChatGPT, LLAMA) a partire dal primo vocabolario individuato manualmente e procedere poi ad un ulteriore *fine tuning*. In particolare BERT si rivela molto performante per una serie di task come sentiment analysis, named entity recognition, coinvolgimento testuale (ovvero la predizione della frase successiva), etichettatura dei ruoli semantici, classificazione del testo e coreference resolution. Di tutte queste possibilità ho sfruttato il coinvolgimento testuale, per ottenere i nuovi termini da riutilizzare nell'esplorazione del corpus. Con questo modello linguistico si è potuto infatti procedere ad una nuova generazione di concordanze a partire dall'elenco iniziale e riapplicarne la ricerca allo stesso corpus di testi (riservandomi in un prossimo futuro di estendere il corpus testuale).

In conclusione, ritengo che compilare un elenco di temi e motivi della ecologia letteraria possa contribuire a sviluppare una maggiore consapevolezza ambientale e a promuovere la letteratura ecologica (evidenziando il valore delle opere letterarie che trattano questioni ambientali e contribuendo a diffondere la consapevolezza su questo genere letterario), fornendo un punto di partenza per la ricerca (individuare le aree in cui esiste ancora un vuoto di conoscenza) e la discussione critica (annotazione tematica, esaltazione di aspetti finora sopiti, analisi di orientamenti storico-geografici ancora inesplorati). L'influenza della natura sui vari aspetti della nostra cultura e dei suoi sviluppi non è ancora stata adeguatamente studiata come ci ricorda Lars Langer [12], il cui approccio computazionale all'analisi delle biodiversità ha mostrato quale elevato potenziale ci sia da sfruttare nell'applicazione di questo tipo di metodologie.

## BIBLIOGRAFIA

- [1] Scaffai Niccolò. Letteratura e Ecologia. Carocci. Roma. (2021).
- [2] Rueckert William. «Literature and Ecology: An Experiment in Ecocriticism». Iowa Review 9, no. 1 (1978).
- [3] Meeker Joseph W. The Comedy of Survival: Studies in Literary Ecology. New York. (1972).
- [4] Lemenager Stephanie. Environmental Criticism for the Twenty-First Century. Routledge (2011)
- [5] Blei, D.M. (2012) Probabilistic Topic Models. Communications of the ACM, 55, 77-84. <http://dx.doi.org/10.1145/2133806.2133826>
- [6] Sollors Werner. The Return of Thematic Criticism. Cambridge (MA)/London: Harvard University Press. (1993)
- [7] Ciotti Fabio, Lana Maurizio, Tomasi Francesca. TEI, ontologies, linked open data: Geolat and beyond. Journal of the Text encoding Initiative. (2014)
- [8] Moretti, F. Distant Reading. London: Verso. (2013)
- [9] Moretti, F. Operationalizing: Or, the Function of Measurement in Literary Theory. New Left Review 84: 103-119. (2013).
- [10] Underwood, T. Topic modeling made just simple enough. The Stone and the Shell. (2012). <https://tedunderwood.wordpress.com/2012/04/07/topic-modeling-made-just-simple-enough/>
- [11] Blei, D. Topic modeling and digital humanities. Journal of Digital Humanities, 2.1 (2013).
- [12] Langers, L. et al, The rise and fall of biodiversity in literature: A comprehensive quantification of historical changes in the use of vernacular labels for biological taxa in Western creative literature, «People and Nature». 3 (5): 1093–1109.

# Risorse linguistiche per lo studio dei *Promessi Sposi*

Rachele Sprugnoli<sup>1</sup>, Arianna Redaelli<sup>2</sup>, Marco Sartor<sup>3</sup>

<sup>1</sup> Università di Parma, Italia – rachele.sprugnoli@unipr.it

<sup>2</sup> Università di Parma, Italia – arianna.redaelli@unipr.it

<sup>3</sup> Università di Parma, Italia – marco.sartor@unipr.it

## ABSTRACT

In questo contributo vengono presentate le attività finalizzate allo sviluppo di nuove risorse linguistiche che possano supportare l'analisi dei *Promessi Sposi* sotto vari aspetti, combinando metodi e strumenti della linguistica computazionale e dell'informatica umanistica. Queste risorse linguistiche sono progettate per avere un impatto positivo in vari campi; potranno, ad esempio, essere utilizzate da ricercatori in ambito umanistico per svolgere analisi linguistiche, da studiosi di *Natural Language Processing* per sviluppare strumenti automatici, da docenti delle scuole superiori per introdurre con metodi digitali i testi manzoniani agli studenti.

## PAROLE CHIAVE

Risorse linguistiche; annotazione linguistica; letteratura italiana; Alessandro Manzoni

## 1. INTRODUZIONE

Una risorsa linguistica può essere definita come materiale linguistico in formato *machine readable*, utilizzabile per studi di ricerca linguistica ma anche per lo sviluppo e la valutazione di strumenti di elaborazione del linguaggio naturale (in inglese: *Natural Language Processing*). Tipici esempi di risorse linguistiche sono dizionari elettronici, lessici e corpora. Tra questi ultimi, numerosi sono quelli interamente costituiti da testi letterari o che includono tale genere testuale. Ad esempio, l'infrastruttura CLARIN dà accesso a 44 corpora letterari di 16 lingue diverse:<sup>1</sup> nessuno di questi, tuttavia, contiene testi italiani. DiaCORIS [2], CODIT [3] e MIDIA [4] sono, invece, corpora diacronici italiani comprendenti generi diversi: le opere di Manzoni sono presenti sia in CODIT che in MIDIA. Queste opere sono disponibili anche nel portale Intratext [5] e nel database LIZ (Letteratura Italiana Zanichelli), mentre il sito CBook contiene vari materiali sui *Promessi Sposi* pensati per scopi didattici, e un'interfaccia di ricerca per esplorare il contenuto del romanzo.<sup>2</sup> I testi manzoniani, quindi, risultano attestati in vari siti e corpora, ma nessuna annotazione manuale e solo parzialmente controllati, mentre le annotazioni possono essere interrogate solo tramite un'interfaccia web.

Ancora più rari sono i testi letterari italiani annotati con altre informazioni linguistiche. Ad esempio, per l'italiano, nell'ambito della *sentiment analysis*, possiamo citare il dataset AriEmozione, in cui i versi di arie d'opera italiane del XVIII secolo sono annotati in base all'emozione trasmessa [6]. Al contrario, a livello internazionale sono numerose le iniziative in questo campo: per una panoramica si veda l'approfondita indagine di Kim e Klinger [7].

Infine, per quanto riguarda i corpora paralleli, il sito CLARIN ne elenca 87:<sup>3</sup> solo 5 di queste risorse contengono testi italiani, ma nessuna include la traduzione delle opere manzoniane.

Rispetto ai precedenti lavori appena brevemente presentati, le attività oggetto di questa proposta, e attualmente in via di sviluppo, intendono far avanzare lo stato dell'arte in merito alle risorse linguistiche dedicate ai testi letterari creando e distribuendo: (i) l'annotazione manuale di lemmi, parti del discorso, tratti morfologici e flessivi in alcuni capitoli tratti dalle diverse redazioni dei *Promessi Sposi* (i.e., *Fermo e Lucia*, *Ventisettana* e *Quarantana*); (ii) nuovi modelli di annotazione morfo-grammaticale addestrati sui capitoli annotati come descritto nel punto precedente; (iii) un corpus parallelo allineato a livello di frase formato dalla Ventisettana e dalla sua traduzione inglese del 1834;<sup>4</sup> (iv) alcuni capitoli della Ventisettana allineati a livello di parola con la corrispondente traduzione inglese del 1834; (v) un dataset di soliloqui e dialoghi tratti dalle diverse redazioni dei *Promessi Sposi* e annotati con le emozioni di base. Lo sviluppo delle suddette risorse è attualmente in corso; una volta realizzate, verranno rese disponibili attraverso il repository ILC4CLARIN di CLARIN-IT.<sup>5</sup>

<sup>1</sup> <https://www.clarin.eu/resource-families/literary-corpora>

<sup>2</sup> [https://cbook.it/l\\_promessi\\_sposi/chaptersList.page](https://cbook.it/l_promessi_sposi/chaptersList.page)

<sup>3</sup> <https://www.clarin.eu/resource-families/parallel-corpora>

<sup>4</sup> <https://www.gutenberg.org/ebooks/35155>

<sup>5</sup> <https://ilc4clarin.ilc.cnr.it/>



## 2. METODOLOGIA

Per l'annotazione morfo-grammaticale vengono seguite le linee guida dell'iniziativa Universal Dependencies (UD) [8] e gli annotatori operano correggendo l'output del modello ISDT di UDpipe (versione 2.5) su un foglio di calcolo.<sup>6</sup> Un'analisi preliminare del capitolo XXIII ha evidenziato la necessità di aggiungere alcuni tratti attualmente non utilizzati nelle *treebank* italiane al fine di rendere esplicite delle informazioni linguistiche significative per i testi di riferimento, come la presenza di forme arcaiche, nomi alterati, apocopi. Questi tratti vengono segnalati nella decima colonna del formato CoNLL-U in quanto non ufficialmente attestati in UD. Come si può notare nell'esempio mostrato nella figura 1, si è scelto di lemmatizzare indicando il lemma a noi più contemporaneo ma aggiungendo il lemma arcaico nella decima colonna.

1	«	«	PUNCT	FB	–	–	–	–	–	–
2	Ah	ah	INTJ	I	–	–	–	–	–	–
3	,	,	PUNCT	FF	–	–	–	–	–	–
4	non	non	ADV	BN	PronType=Neg	–	–	–	–	–
5	perdiam	perdere	VERB	V	Mood=Ind Number=Plur Person=1 Tense=Pres VerbForm=Fin	–	–	–	Variant=Apoc	–
6	tempo	tempo	NOUN	S	Gender=Masc Number=Sing	–	–	–	–	–
7	!	!	PUNCT	FS	–	–	–	–	–	–
8	»	»	PUNCT	FB	–	–	–	–	–	–
9	sclamò	esclamare	VERB	V	Mood=Ind Number=Sing Person=3 Tense=Past VerbForm=Fin	–	–	–	sclamare Style=Arch	–
10	Federigo	Federigo	PROPN	SP	–	–	–	–	–	–
11	ansante	ansante	ADJ	A	Number=Sing	–	–	–	–	–

Figura 1. Esempio di annotazione in formato CoNLL-U, frase tratta dalla Ventisettesima.

L'allineamento a livello di frase viene invece eseguito in modo semi-automatico: il software LF Aligner<sup>7</sup> viene applicato capitolo per capitolo e l'output viene corretto manualmente. Per quanto riguarda l'allineamento a livello di parola, è stata adottata l'interfaccia Ugarit [9], ma in futuro verranno valutati degli strumenti automatici, come *fast\_align*<sup>8</sup> e *AWESOME*<sup>9</sup>, per verificare la fattibilità di un approccio non manuale.

Infine, per l'annotazione manuale delle emozioni si prendono in considerazione le otto emozioni di base di Plutchik [10] (i.e., gioia, tristezza, paura, rabbia, fiducia, disgusto, sorpresa, aspettativa), che vengono assegnate a livello di frase. Le annotazioni risultanti saranno utilizzate per valutare gli strumenti automatici esistenti per il rilevamento delle emozioni.

## 3. RINGRAZIAMENTI

La pubblicazione è stata realizzata da ricercatrice con contratto di ricerca cofinanziato dall'Unione europea - PON Ricerca e Innovazione 2014-2020 ai sensi dell'art. 24, comma 3, lett. a), della Legge 30 dicembre 2010, n. 240 e s.m.i. e del D.M. 10 agosto 2021 n. 1062, progetto "Integrating Digital Archives (IDA)". Questa ricerca è stata anche finanziata dall'Università degli Studi di Parma attraverso l'azione Bando di Ateneo 2022 per la ricerca co-finanziata dal MUR-Ministero dell'Università e della Ricerca - D.M. 737/2021 - PNR - PNRR - NextGenerationEU.

## BIBLIOGRAFIA

- [1] Godfrey, John J., e Antonio Zampolli. «Language resources». In *Survey of the state of the art in human language technology*, a cura di Giovanni Battista Varile e Antonio Zampolli, 381-84. *Linguistica computazionale*, Cambridge (MA); Pisa: Cambridge University Press; Giardini editori e stampatori, 1997.
- [2] Onelli, Corinna, Domenico Proietti, Corrado Seidenari, e Fabio Tamburini. «The DiaCORIS project: a diachronic corpus of written Italian». In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*, maggio 2006, Genova. European Language Resources Association (ELRA), 1212-15.
- [3] Micheli, M. Silvia. «CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field». *Corpus*, fasc. 23 (2022).
- [4] D'Achille, Paolo, e Maria Grossmann, a c. di. *Per la storia della formazione delle parole in italiano: un nuovo corpus in rete (MIDIA) e nuove prospettive di studio*. Quaderni della Rassegna 125. Firenze: Franco Cesati editore, 2017.
- [5] Mastidoro, Nicola. «The IntraText Digital Library: XML-Driven Online Library Based on High Accessibility, Lexical Hypertextualization and Scholarly Accuracy in Philological / Textual Notations». In *Research and Advanced Technology for Digital Libraries*, a cura di Maristella Agosti e Costantino Thanos, 2458: 659-659. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002.
- [6] Zhang, Shibingfeng, Francesco Fericola, Federico Garcea, Paolo Bonora, e Alberto Barrón-Cedeño. «AriEmozione 2.0: Identifying Emotions in Opera Verses and Arias». *Italian Journal of Computational Linguistics* 8, fasc. 2 (2022).

<sup>6</sup> <https://ufal.mff.cuni.cz/udpipe>

<sup>7</sup> <https://sourceforge.net/projects/aligner/>

<sup>8</sup> [https://github.com/clab/fast\\_align](https://github.com/clab/fast_align)

<sup>9</sup> <https://github.com/neulab/awesome-align>

- [7] Kim, Evgeny, e Roman Klinger. «A Survey on Sentiment and Emotion Analysis for Computational Literary Studies». *Zeitschrift für digitale Geisteswissenschaften*, 2019.
- [8] Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre, e Daniel Zeman. «Universal Dependencies». *Computational Linguistics* 47, fasc. 2 (2021): 255-308.
- [9] Palladino, Chiara, Maryam Foradi, e Tariq Yousef. «Translation Alignment for Historical Language Learning: a Case Study». *Digital Humanities Quarterly* 15, fasc. 3 (2021).
- [10] Plutchik, Robert. «A general psychoevolutionary theory of emotion». In *Theories of emotion*, 3-33. Cambridge (MA): Academic Press, 1980.

## **4. Preservazione della memoria e del patrimonio digitale**

# Uso e riuso di dati per la valorizzazione del patrimonio culturale italiano online: buone pratiche per l'interoperabilità

Chiara Aiola<sup>1</sup>, Giulia Fabbris<sup>2</sup>

<sup>1</sup> Net7, Italia – aiola@netseven.it

<sup>2</sup> Università Ca' Foscari Venezia, Italia – giulia.fabbris@unive.it

## ABSTRACT

Con questo poster si intende presentare una riflessione sulla necessità di seguire standard e buone pratiche per lo sviluppo di progetti e l'inserimento di oggetti nell'ecosistema digitale al fine di creare risorse interoperabili in un'ottica collaborativa e di riuso.

## PAROLE CHIAVE

Interoperabilità; riuso di dati; patrimonio culturale digitale; piano nazionale di digitalizzazione; principi FAIR

## 1. INTRODUZIONE

Nonostante la fama che ormai vantano a livello internazionale, i principi [FAIR](#) non sembrano ancora alla base di molti progetti digitali. La scarsa osservanza di standard e linee guida, l'utilizzo di formati diversi e il limitato libero accesso alle risorse digitali prodotte sono grossi impedimenti per lo sviluppo di un vero e proprio ecosistema digitale del patrimonio culturale. Certamente esistono standard informatici così come linee guida per la codifica di oggetti digitali (TEI, DCMI), ma manca la volontà da parte di molti studiosi di integrare questi strumenti o “buone pratiche” nei loro progetti. È infatti necessario che le risorse prodotte siano interoperabili, poiché “[l]’ambiente digitale trova la propria essenza costitutiva nelle relazioni, ovvero nella possibilità di generare e rigenerare connessioni reciproche tra le informazioni, facilitando la produzione di nuovi significati” (PND:17). Al contrario, ogni progetto costituirebbe un’entità a sé stante senza possibilità di integrazione e di riuso.

Anche la mancanza di un’infrastruttura informatica specificamente dedicata alla conservazione dei dati nell’ambito delle Digital Humanities spesso costituisce un ostacolo al reperimento dei dati. Si assiste spesso a un proliferare di progetti e piattaforme legati a specifici progetti che non espongono API per la condivisione dei dati (o se esistono non sono documentate), oppure si sceglie di utilizzare strumenti che non nascono propriamente per la conservazione di dati ma che consentono lo storage gratuito e persistente del materiale caricato, come repository su GitHub. Questa scelta da un lato tutela i ricercatori dal rischio di perdere il proprio lavoro e dall’impossibilità di sostenere costi di hosting, consente il libero accesso alle risorse, anche attraverso il sistema di API fornite dalla piattaforma, ma dall’altro non permette di avere strumenti specifici per strutturare i metadati o i file del progetto, rendendo quindi problematico un processo di automatizzazione dell’interpretazione dei dati (cf. sotto la creazione di un’infrastruttura del patrimonio culturale promossa dal Ministero della cultura nell’ambito del [Piano Nazionale di Digitalizzazione](#)).

Un altro aspetto che è da tenere in grande considerazione è il concetto di obsolescenza. Le risorse analogiche sono deteriorabili, per questo è bene che vengano digitalizzate. Ma le risorse digitali o digitalizzate non sono meno soggette allo stesso destino, poiché la tecnologia è in continua evoluzione e linguaggi, software e hardware diventano obsoleti molto velocemente. Per andare incontro alla voce A2 dei principi FAIR<sup>1</sup>, i dati e i metadati devono poter essere recuperabili anche quando le piattaforme o i sistemi informatici utilizzati per la loro raccolta e visualizzazione vengono dismessi o le tecnologie usate diventano obsolete, e per fare ciò è necessario adottare modelli condivisi e controllati. Quanto evidenziato finora rientra nei primi tre punti di debolezza identificati dal Piano Nazionale di Digitalizzazione (PND) del Patrimonio Culturale circa lo stato attuale dell’ecosistema digitale italiano per il patrimonio culturale (PND:5).

Con questo poster si intende quindi avanzare una riflessione sulla necessità di uniformità dei dati e sull’utilizzo di buone pratiche, elaborata a partire dai risultati parziali di un progetto dottorale [PON](#) in corso di svolgimento. Il progetto, in linea con il PND, ha come scopo la creazione di un’applicazione per la valorizzazione del patrimonio culturale italiano e si basa sul riuso delle risorse già presenti nel web e disseminate in più piattaforme eterogenee (cf. Immagine 1). Il beneficio

---

<sup>1</sup> “Metadata are accessible, even when the data are no longer available”.

primario è quello di preservare il patrimonio culturale messo già a disposizione in rete da altre istituzioni e individui e renderlo più fruibile per una platea diversificata di utenti.

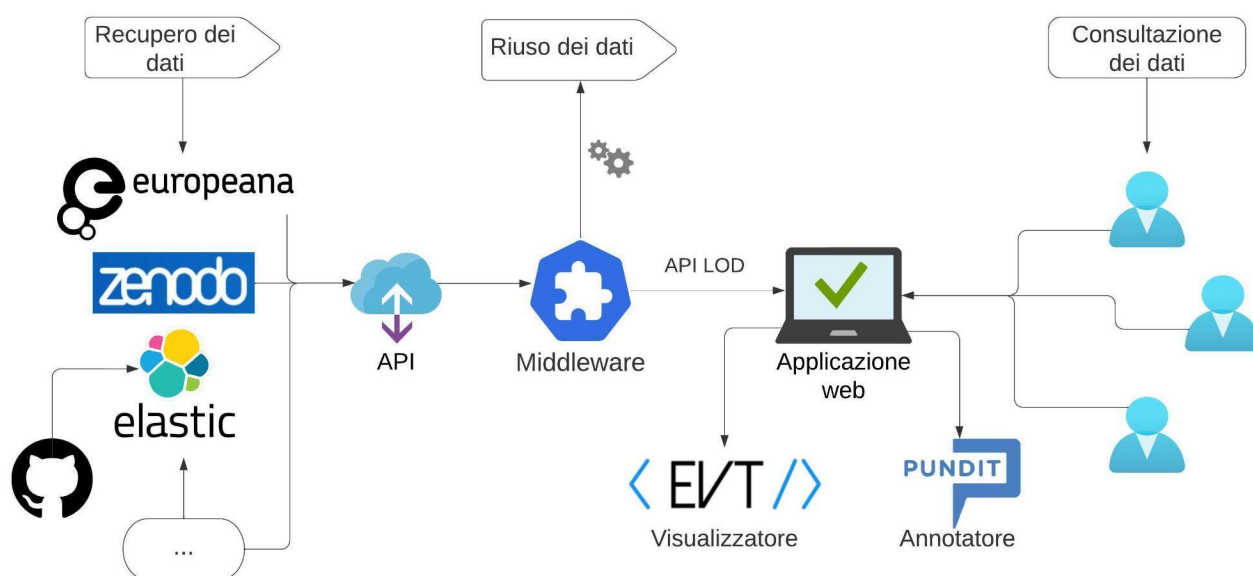


Figura 1. Workflow.

## 2. STATO DELL'ARTE

Negli ultimi anni, e in particolare a seguito della pandemia, c'è stata una forte spinta verso la digitalizzazione delle risorse, sia a livello nazionale che europeo. Con l'aumento di fondi e finanziamenti indirizzati allo sviluppo digitale e una maggior attenzione alla preservazione del patrimonio culturale, diversi aggregatori sono sorti. A livello regionale possiamo citare a titolo esemplificativo e non esaustivo i portali [Puglia Digital Library](#), [Cultura Campania](#) e [Mèmorà](#) (Piemonte), mentre sul piano nazionale è opportuno fare riferimento a [CulturaItalia](#). Questo strumento è di grande rilevanza per il progetto oggetto di questo poster, poiché è il principale aggregatore italiano di dati per [Europeana](#), molti dei quali confluiranno nell'app. È inoltre un ottimo esempio di *FAIRness*, poiché recupera metadati esistenti tramite un Application Profile basato sulle linee guida della DCMI e li restituisce a Europeana in un ulteriore formato standard, EDM, basato anch'esso su capisaldi dell'interoperabilità quali OAI ORE, DC, SKOS e CIDOC-RM. (cf. <https://pro.europeana.eu/page/edm-documentation> e [3, 7])

È inoltre recentissima la pubblicazione della versione 1.1 del PND, tramite il quale il Ministero della cultura intende promuovere e, fondamentale, organizzare la trasformazione digitale dei beni culturali in Italia, andando a supportare gli enti che operano in questo settore. Un importante obiettivo esposto nel Piano è la creazione di un'infrastruttura software del patrimonio culturale per organizzare le risorse del patrimonio culturale e dei metadati in maniera corretta, affidabile, sicura ed efficiente (PND:35). Questo spazio è pensato per essere un supporto integrativo nel quale far confluire le diverse risorse provenienti dai vari domini dell'ecosistema per agevolarne il confronto.

Un altro aspetto molto importante è la documentazione pubblicata a disposizione dei fruitori dei vari siti e progetti (cf. [11]). Nella loro indagine, Franzini et al. 2019 cercano di capire quali sono le aspettative degli utenti e gli usi più comuni circa le edizioni digitali di testi. Sebbene il progetto in sviluppo dalla proponente si occupi di diversi tipi di risorse del patrimonio, alcuni risultati ottenuti dallo studio appena citato sono perfettamente estendibili a progetti digitali di respiro più ampio. In particolare, alla domanda 8) *How important is detailed editorial and technical documentation?* risponde il 69% dei partecipanti con *important* o *very important*. Gli autori commentano dicendo che “[d]ocumenting the process of creation behind any type of project serves to communicate development and quality, and to give appropriate context. With documentation, creators [...] facilitate the reuse of a resource. One of the basic principles and assumptions of research is *reproducibility* or, in other words, the ability of one researcher to take the work of another researcher follow the original researcher’s pathway and arrive at the same results. Reproducibility is key to research acceptance and validation” ([9]:10-11). La riproducibilità e quindi la presenza di documentazione è stato un aspetto chiave per la scelta dei progetti da integrare nell'applicazione che si sta sviluppando. Infatti, diversi lavori che erano stati inizialmente presi in considerazione per l'integrazione dei dati sono stati poi esclusi poiché non esponevano le modalità di recupero di questi.

Infine, è necessario citare la *reusability* dei metadati, poiché è un aspetto cruciale per far sì che la risorsa digitale sia preservabile nel tempo. Barbuti 2021 propone un'espansione della *R* dei principi FAIR in *reusable, relevant, reliable e resilient*, ritenendo che questo ampliamento, con particolare riferimento ai metadati descrittivi, permetta di rendere le risorse digitali sostenibili e permanenti se presentano una corretta proporzione tra qualità e quantità di contenuto. Su questo concetto in particolare si fonda il presente progetto dottorale, poiché è previsto esclusivamente il riuso di dati e strumenti già presenti nell'ecosistema digitale, modellati in base alle esigenze specifiche della ricerca in corso al fine di offrire una fruizione alternativa delle risorse.

### 3. OBIETTIVI

Lo scopo di questo poster sarà quindi di mettere in luce alcune delle soluzioni FAIR previste per lo sviluppo del progetto dottorale della proponente.

Il progetto nello specifico si propone di raccogliere in un'unica piattaforma risorse relative al patrimonio culturale italiano e integrare tra loro diversi strumenti di visualizzazione e ricerca. Più in generale, questo poster vuole essere esemplificativo di un approccio in linea con i principi FAIR per la creazione di un nuovo portale, i quali vengono osservati sin dal concepimento del lavoro. Infatti, non è sufficiente che essi vengano integrati in fasi successive a fini, per esempio, di pubblicazione, poiché tale approccio tende a indebolire il potenziale di riuso e integrazione dei dati. È necessario quindi che la ricerca aspiri continuamente alla *FAIRness* e ai fondamenti della gestione e condivisione dei dati durante tutto il processo di concepimento e realizzazione di un progetto. [6] A questo proposito, i cardini su cui si fonda questa ricerca sono:

- il riutilizzo delle risorse già disponibili e riutilizzabili, selezionate in base ai temi della ricerca;
- la scelta di strumenti *open source* per la visualizzazione e rielaborazione dei dati già esistenti;
- l'integrazione tra risorse e applicazioni;
- l'arricchimento dei contenuti.

Le domande implicite a cui questo progetto cerca di rispondere sono: è davvero possibile con gli strumenti attualmente esistenti realizzare un progetto basandosi esclusivamente su risorse e dati esistenti, sviluppando al massimo un "connettore" per mettere in comunicazione il tutto e aggregarlo in un unico ambiente che risulti comunque di semplice utilizzo per un utente specialista e non? Che sforzo richiede questa operazione e che tipo di conoscenze e competenze è stato necessario possedere?

Come accennato sopra, le risorse che saranno visualizzabili tramite questa applicazione verranno recuperate da aggregatori online tramite API. Per ora sono state sviluppate per Europeana, Zenodo, GitHub e si sta lavorando sulla connessione a un endpoint SPARQL sul patrimonio culturale di origine germanica in Veneto, anch'esso in sviluppo con un progetto PON. Viene estratto un insieme di metadati che vengono ricondotti al formato di metadati DCMI per uniformarne la consultazione così da creare dati ricercabili e interoperabili.

Nella piattaforma saranno integrati un software di visualizzazione e uno strumento per le annotazioni. Per implementare la prima funzionalità è stato scelto [Edition Visualization Technology](#), il quale dovrà essere adattato ai bisogni specifici del progetto per poter permettere all'applicazione di recuperare risorse esterne alla build. Questo strumento è ampiamente usato nei progetti di filologia digitale poiché è stato concepito per essere *open source* (il codice è scaricabile liberamente da [GitHub](#)), basato sugli standard del web e molto flessibile per essere adattato, appunto, alle esigenze di progetti differenti. In particolare, riceve come input file codificati in XML/TEI, è implementato nel framework [Angular](#) e supporta la visualizzazione di immagini [IIIF](#) (cf. [4], [8]).

Le annotazioni saranno invece possibili tramite l'installazione di un'estensione, [Pundit](#)<sup>2</sup>, sviluppata dall'azienda partner del progetto [Net7](#), la quale è disponibile anche sull'[EOSC Marketplace](#). Anch'essa è rilasciata con licenza pubblica, può essere usata in ottica collaborativa e offre funzionalità quali evidenziazione, commenti, tag e annotazioni semantiche. Quest'ultime possono essere annotazioni in senso stretto (metadati) oppure grafi risultanti dai metadati e le relazioni nelle varie risorse web che l'utente ha creato con Pundit. È presente anche una dashboard che permette di gestire tutte le annotazioni che possono essere anche filtrate ed esportate in vari formati standard e/o aperti ([5]). Allo stesso modo, l'interfaccia utente sarà simile a quella visualizzabile nella [demo](#) di Muruca, sempre sviluppata dalla suddetta azienda. In particolare, questo strumento permetterà di organizzare le risorse in collezioni e creare percorsi ([1], [10]).

---

<sup>2</sup> È disponibile anche un video introduttivo: <https://youtu.be/sfGS6aYsV3s>

Un primo tentativo di risposta alle domande sopra avanzate può essere quindi che un progetto di questo tipo è implementabile perché gli strumenti e le risorse FAIR esistono e possono essere integrati per dar loro valore aggiunto. Tuttavia, sarà necessario adattarli ai bisogni specifici della ricerca e bisognerà comunque fare i conti con i “limiti” e la qualità più o meno adeguata di ciò che è presente in rete. La risposta implicita alla seconda domanda è che oltre alla collaborazione tra esperti di discipline umanistiche e sviluppatori informatici occorre sviluppare una sensibilità comune alle varie questioni di riusabilità dei dati e del codice. La multidisciplinarietà del settore delle Digital Humanities consiste anche in questo: un affiancamento di esperti di vari ambiti ma che condividono una conoscenza comune di buone pratiche e metodologie.

## BIBLIOGRAFIA

- [1] Aiola, Chiara, Giulio Andreini, Francesca Di Donato, e Tiziana Lombardo. “Sharing knowledge digitally, the Mura case study.” In *AIUCD 2021 – DH per la società: e-guaglianza, partecipazione, diritti e valori nell’era digitale. Raccolta degli abstract estesi della 10° conferenza nazionale 2021*, (2021): 428-431. <https://aiucd2021.labcd.unipi.it/book-of-abstracts/>
- [2] Barbuti, Nicola. “Thinking digital libraries for preservation as digital cultural heritage: by R to R4 facet of FAIR principles.” *International Journal on Digital Libraries* 22, (2021): 309–318. <https://doi.org/10.1007/s00799-020-00291-7>
- [3] Buonazia, Irene, Maria Emilia Masci. “Il PICO Application Profile. Un Dublin Core Application Profile per il Portale della Cultura Italiana.” In *Interoperabilità di contenuti e servizi digitali: metadati, standard e linee guida*. (2007): 45-51. <https://hdl.handle.net/11384/13790>
- [4] Cacioli, Giulia, Giacomo Cerretini, 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 - Digital cultures. Intersections: philosophy, arts, media. Proceedings of the 11th national conference, Lecce, 1-3 June 2022*, Fabio Ciraci, Giulia Miglietta e Carola Gatto (ed.). (2022): 212-217. <http://amsacta.unibo.it/6848/>
- [5] De Santis, Luca, Duccio Breschi, Niccolò Cardelli, Edgar Gomez, Massimiliano Pardini, Marco Zizi, Maxime Bouillard, Simone Kopeinik e Laurent Capelli. “TRIPLE Deliverable: D5.5 Report on the Open Annotation Tool (Draft).” Zenodo. (2021). <https://doi.org/10.5281/zenodo.5653219>
- [6] Dempsey, William, Ian Foster, Scott Fraser e Carl Kesselman. “Sharing Begins at Home: How Continuous and Ubiquitous FAIRness Can Enhance Research Productivity and Data Reuse.” In *Harvard data science review* 4, no. 3 (2022). <https://doi.org/10.1162/99608f92.44d21b86>
- [7] Di Giorgio, Sara. “Culturaitalia, the Italian National Content Aggregator in Europeana.” In *Procedia Computer Science* 38, (2014): 40-43. <https://doi.org/10.1016/j.procs.2014.10.008>
- [8] Di Pietro, Chiara, Chiara Martignano e Roberto Rosselli Del Turco. “Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo.” *Umanistica Digitale*, no. 7 (2019). <http://doi.org/10.6092/issn.2532-8816/9322>
- [9] Franzini, Greta, Melissa Terras e Simon Mahony. “Digital Editions of Text: Surveying User Requirements in the Digital Humanities.” *ACM Journal on Computing and Cultural Heritage* 12, no. 1 (2019). <https://doi.org/10.1145/3230671>
- [10] Lombardo, Tiziana e Chiara Aiola. “Storytelling through Digital Scholarly Editions.” Zenodo. (2022). <https://doi.org/10.5281/zenodo.6594346>
- [11] Warwick, Claire, Isabel Galina, Jon Rimmer, Melissa Terras, Ann Blandford, Jeremy Gow e George Buchanan. “Documentation and the users of digital resources in the humanities.” *Journal of Documentation* 65, no. 1 (2009): 33-57. Emerald Group Publishing Ltd. <https://www.doi.org/10.1108/00220410910926112>



# “Dalla civiltà dell'archivio alla civiltà dell'algorithmo?”

Alessandro Alfier<sup>1</sup>

<sup>1</sup> Scuola di dottorato in scienze documentarie, linguistiche e letterarie della Sapienza Università di Roma, Italia –  
alessandro.alfier@uniroma1.it

## PAROLE CHIAVE

Autenticità; blockchain; conservazione digitale; fiducia; OAIS

## 1. OBIETTIVO DELLA PROPOSTA

Nell'ambito della conservazione della memoria e del patrimonio culturale digitale va delineandosi un nuovo paradigma che, esplicitamente, intende sostituirsi a quello che potremmo definire il modello consolidato derivato dalla tradizione. Quest'ultimo può essere definito come un "modello archivistico-custodiale": incentrato dunque sulle funzioni di preservazione garantite dall'archivio, inteso come sistema e organizzazione che fanno capo a un soggetto votato al perseguimento di finalità conservative a favore di ciò che è stato socialmente individuato come patrimonio culturale. Tale modello, non a caso, è stato integralmente ripreso dal più rilevante e influente *reference model* in ambito di preservazione digitale: l'*Open Archival Information System* (OAIS), elaborato dal *Consultative Committee for Space Data Systems*<sup>1</sup>. Il rapporto che sussiste tra questo standard *de facto* e il tradizionale modello di custodia è così stretto che l'OAIS può essere correttamente interpretato anche come il tentativo di tradurre quel consolidato paradigma dall'originario scenario analogico al nuovo scenario digitale. Lo standard infatti non prescrive specifici requisiti di natura tecnologica, né propone delle precise soluzioni applicative: coerentemente con la propria natura di *reference model*, esso in realtà esplicita e formalizza i concetti e le funzioni basilari a cui un custode deve attenersi, per poter preservare a lungo termine e attraverso un sistema e un'organizzazione d'archivio un patrimonio culturale digitale inteso in senso ampio e che può pertanto comprendere beni archivistici, beni bibliografici e qualsiasi collezione di oggetti digitali riconosciuti come dimensione memoriale. Tanto è vero che se si scorrono le definizioni presenti nel glossario che correda OAIS<sup>2</sup> si ritrovano, sostanzialmente, tutti gli aspetti essenziali che connotano il tradizionale modello di custodia, aggiornati al nuovo scenario digitale:

- l'*Open Archival Information System* (OAIS) è definito come «an archive, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a designated community»<sup>3</sup>. In questa definizione si ritrova riproposto il modello consolidato di custodia, incentrato sulla nozione d'archivio come dimensione organizzativa e di sistema, posta sotto la responsabilità di un soggetto che opera in favore di una comunità di utenti;
- a sua volta il concetto di *Archive* è definito come «an organization that intends to preserve information for access and use by a designated community»<sup>4</sup>, intesa quest'ultima come «an identified group of potential consumers who should be able to understand a particular set of information. The designated community may be composed of multiple user communities. A designated community is defined by the archive and this definition may change over time»<sup>5</sup>. In tali definizioni si ritrova la visione del paradigma tradizionale, secondo cui la custodia consiste concretamente in quello spazio operativo che vive delle relazioni intercorrenti tra il ruolo del custode e quello degli utenti;
- si precisa inoltre in che cosa consista la finalità di conservazione a lungo termine, attraverso il concetto di *Long Term Preservation*: «the act of maintaining information, independently understandable by a designated community, and with evidence supporting its authenticity, over the long term»<sup>6</sup>. Tale definizione ripropone

<sup>1</sup> Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS). Recommended practice CCSDS 650.0-M-2. Magenta Book, 2nd ed. (Washington: Consultative Committee for Space Data Systems, 2012), <https://public.ccsds.org/pubs/650x0m2.pdf> (consultato il 5 maggio 2023).

<sup>2</sup> Consultative Committee for Space Data Systems, 1.8-1.16.

<sup>3</sup> Con riferimento alla stessa voce del glossario, nel testo di OAIS si precisa che «the term "open" in OAIS is used to imply that this Recommendation and future related Recommendations and standards are developed in open forums, and it does not imply that access to the archive is unrestricted» (Consultative Committee for Space Data Systems, 1.13).

<sup>4</sup> Consultative Committee for Space Data Systems, 1.9.

<sup>5</sup> Consultative Committee for Space Data Systems, 1.11.

<sup>6</sup> Consultative Committee for Space Data Systems, 1.13.

la prospettiva del modello tradizionale, secondo cui la custodia trova la sua ragion d'essere nel preservare sul lungo periodo - potenzialmente a tempo indefinito - i patrimoni oggetto di custodia, garantendo la loro inalterabilità sostanziale e dunque l'autenticità dei significativi e dei valori in essi iscritti.

È ben vero che, in questa riproposizione del classico modello di custodia, OAIS introduce anche degli elementi di novità, al fine di attualizzare quel paradigma al nuovo scenario digitale. Tali innovazioni sono rintracciabili su due piani precisi. In primo luogo attraverso la definizione di *Designated Community* poco sopra riportata, che introduce nel paradigma di custodia una componente fortemente dinamica: quella stessa definizione precisa infatti che l'archivio, inteso come organizzazione e sistema finalizzati alla conservazione del patrimonio culturale digitale, non è definito una volta per tutte, ma deve essere capace di evolvere coerentemente con il mutare, nel tempo, delle esigenze della comunità designata. In questo modo i bisogni degli utenti, visti in una prospettiva diacronica, diventano una variabile importante nella definizione delle strategie conservative, facendo definitivamente venir meno quella logica auto-referenziale della tradizione, in ragione della quale le stesse strategie conservative erano invece per lo più formalizzate sulla base delle metodologie disciplinari di trattamento dei materiali custoditi e sulla base delle esigenze dei custodi, relegando gli utenti a un mero ruolo di fruitori passivi<sup>7</sup>. A questo proposito si potrebbe anche sostenere che, attraverso il concetto di *Designated Community*, il modello di custodia si appropria dell'approccio *user-centered design*, che è andato diffondendosi proprio sull'onda dell'espansione dello scenario digitale. Una seconda innovazione introdotta da OAIS, nella sua attualizzazione del paradigma tradizionale, è legata alla definizione di conservazione a lungo termine che, come evidenziato dalla citazione del glossario poco sopra riportata, a sua volta rimanda a due altri concetti chiave: a quello di *Long Term*, inteso come «a period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing designated community, on the information being held in an OAIS. This period extends into the indefinite future»<sup>8</sup>; e a quello di conservazione dell'informazione come *Independently Understandable*, intendendosi con tale espressione «a characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the designated community without having to resort to special resources not widely available»<sup>9</sup>. In questi sforzi definitivi è evidente l'esigenza di declinare il modello tradizionale di custodia in rapporto a due problematiche profondamente avvertite nel nuovo scenario digitale: da un lato quello dell'obsolescenza tecnologica, che inevitabilmente investe il patrimonio culturale custodito nella misura in cui esso si compone di oggetti digitali e che in una prospettiva a lungo termine comporta allora la previsione di una serie di attività a contrasto della stessa obsolescenza tecnologica; dall'altro lato quello della dipendenza da soluzioni applicative e software che possono condizionare, in modo importante, l'intelligibilità da parte degli utenti degli oggetti digitali e che pertanto, in una prospettiva a lungo termine, richiede di essere adeguatamente gestita, per evitare che la piena fruizione del patrimonio culturale digitale conservato sia soggetta a restrizioni dovute al contesto tecnologico. Da notare che entrambe queste due innovazioni evidenziano un legame con il concetto di dinamicità: nel primo caso si è dinnanzi a una novità che comporta l'immissione di dinamicità nel consolidato paradigma di custodia, attraverso la considerazione della variabilità dei bisogni dell'utenza; nel secondo caso si è invece di fronte a una novità con cui si pretende di supportare il tradizionale modello conservativo nel suo sforzo di far fronte alla dinamicità di per sé insita nel nuovo orizzonte tecnologico.

In questo intento di attualizzazione della dimensione di custodia presente nel *reference model* di OAIS e realizzato mediante la ricerca di un ideale punto d'equilibrio tra il piano della stabilità e quello della dinamicità, vi è un concetto che risulta formulato in modo piuttosto generico nel testo dello stesso standard: quello d'autenticità. Va ricordato che tale nozione è strettamente legata al modello consolidato di preservazione, in ragione del fatto che il custode è tradizionalmente chiamato

---

<sup>7</sup> Il concetto di *Designated Community* dischiude non solo a una visione definitivamente incentrata sui bisogni degli utenti, ma anche a una contestuale crescita del ruolo di mediazione tradizionalmente già assegnato ai conservatori. Nel modello consolidato i custodi sono chiamati ad operare come mediatori - nell'ambito della relazione d'uso tra gli utenti e il patrimonio culturale conservato - in quanto la funzione custodiale non abbraccia solo azioni di preservazione in senso stretto, ma anche quelle per la predisposizione degli strumenti necessari agli utenti per poter far uso dei beni culturali conservati. Tale ruolo di mediazione risulta, ad esempio, particolarmente penetrante nella fattispecie della conservazione dei beni archivistici: nella figura dell'archivista custode si riconosce infatti un vero e proprio mediatore, in quanto egli deve individuare le tecniche di ordinamento e gli strumenti di ricerca più idonei per mettere a disposizione degli utenti una memoria-fonte che originariamente è stata però prodotta con ben altre finalità, segnatamente come memoria-autodocumentazione al servizio di persone e organizzazioni (Isabella Zanni Rosiello, *Archivi e memoria storica* (Bologna: il Mulino, 1987)), rendendosi così protagonista di una complessa operazione intellettuale di "traduzione". Ebbene il concetto di *Designated Community* accentua ancora più tale dimensione di mediazione, perché ora i custodi non sono solo chiamati alla predisposizione degli strumenti d'uso del patrimonio culturale conservato sulla base dei tradizionali criteri tecnico-disciplinari, ma ancor prima sono esplicitamente incaricati dell'interpretazione ad ampio raggio dei bisogni dell'utenza, sulla base dell'assunzione che gli stessi possono variare nel corso del tempo.

<sup>8</sup> Consultative Committee for Space Data Systems, 1.12.

<sup>9</sup> Consultative Committee for Space Data Systems, 1.12.

a conservare il patrimonio culturale non in un modo qualsiasi, ma secondo requisiti atti a proteggere la credibilità dei beni culturali detenuti, così da assicurare che essi mantengano inalterati nel tempo i significati e i valori originariamente acquisiti al momento della loro primigenia produzione. Tale concetto in OAIS è espresso attraverso il termine di *Authenticity*, che nel glossario è definita come «the degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence»<sup>10</sup>. Si tratta di una definizione che, con tutta probabilità, ha ricevuto una formulazione generica sulla base di una precisa esigenza. La nozione di autenticità presenta infatti una notevole variabilità semantica: innanzitutto sul piano diacronico, perché come per ogni concetto soggetto all'evoluzione storica anch'esso è andando assumendo significati diversi in ragione delle differenti epoche; ma soprattutto sul piano sincronico - che è quello che qui più interessa - giacché nei diversi domini che costituiscono quello che oggi possiamo considerare il variegato patrimonio culturale, il concetto d'autenticità si riveste di specifiche declinazioni di significato. OAIS pertanto, coerentemente con la finalità di proporre un *reference model* applicabile trasversalmente alla preservazione di qualsiasi tipologia di oggetto digitale culturale, a prescindere dal dominio disciplinare di appartenenza, ha dovuto far riferimento a una nozione d'autenticità massimamente astratta, sotto la quale fosse possibile sussumere le diverse accezioni del termine in vigore nei diversi settori di quello che noi oggi consideriamo patrimonio culturale. Questa stessa variabilità semantica può essere a sua volta schematizzata ricorrendo a una distinzione basilare. Per alcuni domini del *cultural heritage* il concetto di autenticità tende infatti a identificarsi sostanzialmente con quello di integrità, mentre per altri la nozione d'autenticità è bidimensionale, prevedendo in aggiunta al piano dell'integrità anche quello della veridicità: come se la variabilità semantica di cui si sta discutendo si manifestasse nei termini di un'oscillazione tra un maggior e un minor numero di componenti chiamate, di volta in volta, in causa nella configurazione della dimensione di ciò che consideriamo autentico. Questi processi di differenziazione semantica persistono, naturalmente, anche quando ci si confronta con patrimoni culturali non più analogici, ma digitali. Se ad esempio rivolgiamo l'attenzione al dominio rappresentato dagli archivi storici, oggi sempre più spesso digitali, vi ritroviamo chiaramente una configurazione bidimensionale dell'autenticità: essi infatti si ritengono autentici se la loro custodia li ha protetti da ogni alterazione voluta o accidentale e se essa, al contempo, ha garantito nelle forme dovute la veridicità di cui la stessa documentazione storica è portatrice rispetto ai fatti e agli eventi in essa rappresentati. Solo soddisfacendo entrambi questi requisiti gli archivi storici, tanto analogici quanto digitali, possono essere usati dagli utenti come complessi di fonti credibili e autorevoli, per dar seguito alla ricerca storica<sup>11</sup> o per dimostrare la sussistenza di rapporti giuridici di lunga data. Ciò premesso, diviene allora evidente come il ricorso a OAIS non risolva, *d'emblée*, tutti i problemi relazionati con la conservazione degli oggetti digitali culturali: lo standard infatti, ben lontano da qualsiasi "mito di autosufficienza", fornisce un *reference model* assai generale che deve essere poi opportunamente declinato e dettagliato all'interno del singolo contesto custodiale e in ragione della specificità semantica che il termine di autenticità assume in quello stesso contesto conservativo.

La nozione d'autenticità nella sua configurazione più articolata, che non può prescindere dunque dal valore veridico del patrimonio culturale conservato in quanto rappresentazione credibile e autorevole - e pertanto fidata - della realtà, è il risultato di un lungo processo d'evoluzione storica, il cui momento d'avvio è stato sottoposto a indagine. È stato così sottolineato, sul piano etimologico, il risalente legame tra i termini *authenticum* e *auctoritas* e la loro comune derivazione da *augere*, che esprime l'idea di accrescimento<sup>12</sup>. Secondo quest'interpretazione l'autenticità, nella sua visione più complessa, emergerebbe allora da epoche remote come qualità conferita da un'autorità pubblica fattasi garante della produzione veridica di un insieme di significati che così subiscono un positivo effetto di "accrescimento", in ragione del fatto che come esito di quello stesso processo d'autenticazione essi possono essere considerati come verità fidate dall'intero consesso sociale. Tale interpretazione filologica permette, tra l'altro, d'individuare una delle ragioni che spiegano il perché il modello di custodia si sia tradizionalmente incentrato

---

<sup>10</sup> Consultative Committee for Space Data Systems, 1.9.

<sup>11</sup> Certamente non va dimenticata la riflessione storiografica che nel corso del '900, attraverso le polemiche prese di posizione di studiosi quali Jacques Le Goff, Paul-Michel Foucault e Marc Léopold Benjamin Bloch, ha sottolineato la natura potenzialmente insidiosa della documentazione d'archivio usata come fonte per la ricerca storica, capovolgendo così l'impostazione delle precedenti correnti storiografiche ottocentesche, propense a esaltare un'oggettiva veridicità a priori di quelle stesse fonti. Sul fronte poi della scienza archivistica è recentemente emerso un nuovo approccio, denominato *archival turn*, che riconosce nella conservazione tradizionale degli archivi storici non più il risultato dell'applicazione di pratiche neutrali e imparziali da parte del custode, bensì «historically constructed tools of power relations, deeply embedded in changing social and cultural contexts» (Filippo De Vivo, Andrea Guidi e Alessandro Silvestri, "Archival Transformations in Early Modern European History", *European History Quarterly* 46 (2016): 421), con la conseguenza di mettere in dubbio, anche su questo versante disciplinare, l'oggettiva veridicità a priori delle fonti d'archivio. Va in ogni caso rilevato come questi diversi apporti disciplinari non sembrano negare, in termini assoluti, i valori di veridicità che possono essere opportunamente colti nella documentazione storica, piuttosto sembrano sottolineare come la verità insita nelle fonti d'archivio richieda, per essere individuata, il ricorso a una serie di cautele e avvertenze, così da neutralizzare gli eventuali elementi di ambiguità e distorsione.

<sup>12</sup> Giorgio Costamagna, "I concetti di autenticità e di originalità nella documentazione della Cancelleria genovese nel Medioevo", in *Landesherrliche Kanzleien im Spätmittelalter*, ed. Gabriel Silagi (München: Arbo-Gesellschaft, 1986), 488.

sull'operato di organizzazioni che sono manifestazioni dei poteri pubblici: a evidenziare un antico e intrinseco legame tra la dimensione dell'autentico - nella sua configurazione più articolata - e la sfera delle autorità pubbliche. Riflettendo sulla variabilità semantica della nozione di autenticità si perviene allora a un'ulteriore prospettiva con cui guardare al consolidato modello di custodia, nella sua attualizzazione realizzata dallo standard OAIS: esso emerge infatti, per alcune sue parti, come un vero e proprio paradigma di fiducia, in quanto l'organizzazione e il sistema d'archivio, posti sotto la responsabilità di un custode, sono chiamati a preservare non solo l'integrità, ma anche la veridicità dei significati e dei valori iscritti in alcune tipologie di oggetti culturali digitali, cosicché gli utenti possano fruirne facendo su di essi pieno e completo affidamento.

Proprio il concetto di paradigma di fiducia è oggi però al centro di forti polemiche. Il riferimento inevitabile, a questo proposito, è alla blockchain<sup>13</sup>, spesso evocata come una tecnologia a impatto dirompente rispetto all'odierno quadro economico, politico e sociale. Gli osservatori più avveduti però segnalano come quest'interpretazione debba essere rivista alla luce di una più approfondita chiave di lettura: quella in ragione della quale la blockchain, con la sua architettura computazionale, i suoi protocolli e il suo ricorso alla crittografia, non può solo essere considerata come un'innovazione tecnologica, ma deve essere valutata soprattutto come una potenziale innovazione istituzionale<sup>14</sup>. Nelle intenzioni dei suoi propugnatori, la blockchain dovrebbe infatti fungere da strumento per conseguire uno scopo d'ampia portata: il superamento dei modi tradizionali con cui oggi produciamo e conserviamo quei significati e valori che riconosciamo come autentici e su cui pertanto riponiamo la necessaria fiducia, così da servirne nella costruzione delle relazioni tra individui all'interno del consesso sociale. Modalità tradizionali che nella visione di costoro risentono di gravi limiti, che per l'appunto renderebbero urgente quel superamento: la forte centralizzazione e l'accentuata mediazione in capo alle autorità pubbliche, che si pongono inevitabilmente al centro dei consolidati processi di produzione e conservazione del patrimonio di significati e valori autentici, a tutto discapito di una più diretta partecipazione da parte degli individui e delle loro comunità. La blockchain va dunque interpretata prima di tutto come un'innovazione istituzionale, in quanto essa prospetta certo una soluzione tecnologica, ma in risposta a un problema che è senza dubbio di natura sociale: ridisegnare il concetto di fiducia alla luce delle esigenze della società digitale<sup>15</sup>. E gli intenti che si intendono perseguire con la diffusione della blockchain fanno appello a una risorsa fondamentale: quella rappresentata dalla crittografia pubblica, regolata da standard condivisi e dotata della necessaria maturità, che dunque ha da tempo cessato i panni di tecnica a uso esclusivo degli apparati militari e governativi<sup>16</sup>. È proprio questo forte legame che il paradigma della blockchain manifesta con le tecniche della crittografia a permettere al primo di porsi come modello di messa in discussione dei meccanismi tradizionali di fiducia: l'adozione della crittografia gli porta infatti in dote la capacità di proporre nuovi "regimi di autenticità", basati su parametri algoritmico- matematici.

Il movimento a favore della diffusione della blockchain sembra dunque portare una sfida radicale al consolidato modello di custodia pure aggiornato alla luce di OAIS, in quanto parte di quel tradizionale paradigma di fiducia che è ora messo in discussione alla luce dei meccanismi di potere e di mediazione delle autorità pubbliche: come se la blockchain fosse il cavallo di troia per il definitivo superamento di una civiltà basata sulla custodia d'archivio per preservare significati e valori autentici e inaugurare una nuova era, incentrata invece sull'uso degli algoritmi crittografici per garantire non solo la produzione, ma anche la conservazione dei necessari contenuti credibili. Con la conseguenza di dischiudere un orizzonte in cui la preservazione dei patrimoni culturali digitali non è più affidata a organizzazioni e sistemi iscritti in profili istituzionali caratterizzati anche da una visione gerarchica, ma a processi sostanzialmente tecnologici in mano a comunità di utenti che agiscono secondo una logica orizzontale tra pari. Quanto però è reale questa sfida? La pretesa dei sostenitori della blockchain di annullare le asimmetrie informative e il ruolo

---

<sup>13</sup> La blockchain è definita, in letteratura, come un registro digitale distribuito e condiviso all'interno di una rete pubblica o privata e basato su meccanismi crittografici e di consenso. Centrale in questa nozione è la funzione di registrazione, realizzata per tramite di un database, su informazioni riconducibili a qualsiasi ambito di realtà. Le registrazioni effettuate in una blockchain hanno una precisa organizzazione: sono aggregate in blocchi (*blocks*), che fungono da loro contenitori permanenti e immutabili grazie al ricorso alle tecniche crittografiche. Tali blocchi sono poi tra loro relazionati, a formare una catena che si espande man mano che aumentano le registrazioni effettuate nel corso del tempo: l'estensione della catena coincide con l'interesse della singola blockchain. Il registro è però al contempo distribuito: infatti il database che contiene la catena di blocchi è replicato in copie, dislocate sui diversi nodi della rete che accoglie quella certa blockchain. Per un'analisi approfondita delle caratteristiche tecniche della blockchain si veda Giovanni Michetti, *Introduzione alla blockchain* e Kevin Werback, *The Blockchain and the New Architecture of Trust* (Cambridge: The MIT Press, 2018).

<sup>14</sup> Cosimo Accoto, "Blockchain: innovazione istituzionale", *Aspenia* 96 (2022): 254-261; Cosimo Accoto, *Il mondo ex machina. Cinque brevi lezioni di filosofia dell'automazione* (Milano: Egea, 2019), 63-87.

<sup>15</sup> Giovanni Michetti, *Introduzione alla blockchain. Una guida per archivisti* (Napoli: Editoriale scientifica, 2020), 25; Victoria L. Lemieux, "Blockchain and Public Record Keeping: Of Temples, Prisons, and the (Re)Configuration of Power". *Frontiers in Blockchain*, no. 2 (2019): 1-14, <https://www.frontiersin.org/articles/10.3389/fbloc.2019.00005/full?ref=hackemoon.com> (consultato il 5 maggio 2023).

<sup>16</sup> Jean-François Blanchette, *Burdens of Proof: Cryptographic Culture and Evidence Law in the Age of Electronic Documents*. (Cambridge: The MIT Press, 2012).

centralizzato degli intermediari pubblici, un tempo considerati affidabili, risulta particolarmente critica<sup>17</sup> per due ordini di considerazioni:

- in primo luogo, il modello custodiale attualizzato dallo standard OAIS e incentrato sul concetto d'archivio si dimostra in grado di preservare tanto l'integrità quanto la veridicità del patrimonio culturale digitale conservato, manifestando così la capacità di gestire su di sé una nozione di autenticità complessa e non meramente unidimensionale. Non così invece la blockchain, rispetto a cui verrebbe da porre una serie d'interrogativi: chi assicura che i contenuti registrati nella catena di blocchi siano veridici e non solo integri a seguito del ricorso alle tecniche di crittografia? Chi esercita, in modo trasparente e quindi verificabile, la responsabilità di fornire certezza sull'affidabilità dei contenuti registrati, al di là della loro immodificabilità assicurata dagli algoritmi? La crescita progressiva delle registrazioni che compongono la singola blockchain è infatti sostanzialmente regolata da meccanismi di consenso<sup>18</sup> che, di per sé, non sono titolati a discriminare tra significati e valori veridici e significati e valori non veridici. Pertanto, se si mette in discussione il modello di custodia della tradizione, si perviene a un concetto di autenticità semplificato e inefficace per alcuni contesti, identificabile con la sola nozione di integrità, laddove gli strumenti tradizionali incentrati sulla dimensione della custodia d'archivio alimentano un'autenticità che, ad un tempo, può assumere le forme dell'integrità e della veridicità. Il modello della blockchain sembra pertanto ricondurci a una criticità dirimente: «a poco giova disporre di algoritmi invincibili [...] se poi il collegamento tra mondo reale e mondo crittografico non può essere gestito con la medesima affidabilità»<sup>19</sup>. In questa prospettiva la pretesa dei sostenitori della blockchain di ridurre drasticamente il costo dei processi di gestione della fiducia, eliminando le tradizionali forme gerarchiche di mediazione da parte degli attori pubblici e ricorrendo a un modello orizzontale distribuito *peer to peer* - in cui gli operatori agiscono solo previo consenso della maggioranza e il più delle volte in forma anonima - sembra scontrarsi con la constatazione che quegli stessi costi non sono comprimibili oltre un certo limite, pena l'impoverimento della stessa dimensione di fiducia che pure si vorrebbe rigenerare;
- in secondo luogo, il movimento della blockchain non tiene in considerazione la complessità delle interazioni che nel modello di custodia intercorrono tra i significati e i valori autentici iscritti nel patrimonio culturale digitale conservato e gli usi richiesti dalla comunità di utenti. Tra l'altro, attraverso il concetto di *Designated Community* introdotto da OAIS, l'archivio come dimensione custodiale è chiamato a mutare in funzione dell'evoluzione dei bisogni degli utenti, in un superamento di ogni logica auto-referenziale del custode che è già un primo avvio di quel processo di de-gerarchizzazione e di decentralizzazione dei soggetti pubblici perseguita, con tanta enfasi, dai fautori della blockchain.

## BIBLIOGRAFIA

- [1] Accoto, Cosimo, *Il mondo ex machina. Cinque brevi lezioni di filosofia dell'automazione*. Milano: Egea, 2019.
- [2] Accoto, Cosimo, "Blockchain: innovazione istituzionale". *Aspenia* 96 (2022): 254-261.
- [3] Bechini, Ugo, "Da Berlino a Dublino e Pechino: sulle tracce della blockchain". *Rivista del notariato* 72, no. 6 (2018): 1181- 1190, [https://irp-cdn.multiscreensite.com/0e15f8c7/files/uploaded/Blockchain\\_2019.pdf](https://irp-cdn.multiscreensite.com/0e15f8c7/files/uploaded/Blockchain_2019.pdf)
- [4] Befani, Guido, "Certeza, consenso e certificazioni informatiche: problemi e prospettive di un approccio giuridico al fenomeno delle tecnologie basate sui registri distribuiti". *Il diritto dell'economia* 105, no. 2 (2021): 77-114, [www.ildirittodelleconomia.it/2021/07/23/guido-befani-certeza-consenso-e-certificazioni-informatiche-problemi-e-prospettive-di-un-approccio-giuridico-al-fenomeno-delle-tecnologie-basate-sui-registri-distribuiti/](http://www.ildirittodelleconomia.it/2021/07/23/guido-befani-certeza-consenso-e-certificazioni-informatiche-problemi-e-prospettive-di-un-approccio-giuridico-al-fenomeno-delle-tecnologie-basate-sui-registri-distribuiti/)
- [5] Blanchette, Jean-Fraçois, *Burdens of Proof: Cryptographic Culture and Evidence Law in the Age of Electronic Documents*. Cambridge: The MIT Press, 2012.

---

<sup>17</sup> Guido Befani, "Certeza, consenso e certificazioni informatiche: problemi e prospettive di un approccio giuridico al fenomeno delle tecnologie basate sui registri distribuiti", *Il diritto dell'economia* 105, no. 2 (2021): 77-114, [www.ildirittodelleconomia.it/2021/07/23/guido-befani-certeza-consenso-e-certificazioni-informatiche-problemi-e-prospettive-di-un-approccio-giuridico-al-fenomeno-delle-tecnologie-basate-sui-registri-distribuiti/](http://www.ildirittodelleconomia.it/2021/07/23/guido-befani-certeza-consenso-e-certificazioni-informatiche-problemi-e-prospettive-di-un-approccio-giuridico-al-fenomeno-delle-tecnologie-basate-sui-registri-distribuiti/) (consultato il 5 maggio 2023).

<sup>18</sup> Un aspetto rilevante nel funzionamento della blockchain è rappresentato proprio dal ricorso ai meccanismi di consenso. L'aggiunta di un nuovo blocco alla catena esistente non è infatti lasciata alla libera decisione del singolo nodo della rete, in quanto in questo modo si avrebbe una proliferazione incontrollata dei blocchi stessi, ma è sottoposta al preliminare consenso della maggioranza dei nodi partecipanti.

<sup>19</sup> Ugo Bechini, "Da Berlino a Dublino e Pechino: sulle tracce della blockchain", *Rivista del notariato* 72, no. 6 (2018): 1181-1190, [https://irp-cdn.multiscreensite.com/0e15f8c7/files/uploaded/Blockchain\\_2019.pdf](https://irp-cdn.multiscreensite.com/0e15f8c7/files/uploaded/Blockchain_2019.pdf) (consultato il 5 maggio 2023).

- [6] Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS). Recommended practice CCSDS 650.0-M-2. Magenta Book, 2nd ed. Washington: Consultative Committee for Space Data Systems, 2012, <https://public.ccsds.org/pubs/650x0m2.pdf>
- [7] Costamagna, Giorgio, "I concetti di autenticità e di originalità nella documentazione della Cancelleria genovese nel Medioevo". In *Landesherrliche Kanzleien im Spätmittelalter*, edited by Gabriel Silagi, 485-504. München: Arbo- Gesellschaft, 1986.
- [8] De Vivo, Filippo, Andrea Guidi e Alessandro Silvestri, "Archival Transformations in Early Modern European History". *European History Quarterly* 46 (2016): 421-434.
- [9] Lemieux, Victoria L, "Blockchain and Public Record Keeping: Of Temples, Prisons, and the (Re)Configuration of Power". *Frontiers in Blockchain*, no. 2 (2019): 1-14, <https://www.frontiersin.org/articles/10.3389/fbloc.2019.00005/full?ref=hackemoon.com>
- [10] Michetti, Giovanni, *Introduzione alla blockchain. Una guida per archivisti*. Napoli: Editoriale scientifica, 2020.
- [11] Werback, Kevin, *The Blockchain and the New Architecture of Trust*. Cambridge: The MIT Press, 2018.
- [12] Zanni Rosiello, Isabella, *Archivi e memoria storica*. Bologna: il Mulino, 1987.

# ***Breaking the Canvas***

## **A Serious Game to challenge stereotypes in Visual Culture through Fine Arts**

Selenia Anastasi<sup>1</sup>, Daniele Zolezzi<sup>2</sup>

<sup>1</sup> Università degli Studi di Genova / Language Technology Group, Department of Informatics, Universität Hamburg, Italy and Germany – selenia.anastasi@edu.unige.it

<sup>2</sup> Università degli Studi di Genova, Italy – daniele.zolezzi@gmail.com

### **ABSTRACT**

How are minorities portrayed in visual culture and art history? A large body of literature addresses this question (O’Kelly, 1983). However, relatively little attention has been paid to the potential use of this knowledge as an effective educational tool to raise awareness of social issues in educational contexts. The increasingly widespread digitization of education, the availability of online resources for exploring museum heritage, and circulation of structured and open data could represent an opportunity to experiment with innovative educational projects aimed at promoting themes such as inclusion, equity, and ethical use of new technologies. In this study, we move in this direction by proposing a digital tool to achieve these objectives. To do this, we refer to already established approaches for analysing gender stereotypes in art history and apply them to the educational framework known as a Serious Game. Our contribution is therefore twofold. First, we provide a theoretical and critical reflection on the state of the art of educational digital approaches to the study of artistic heritage from an inclusive perspective, considering the possibility of addressing a relatively young student target. Second, we provide a practical solution by proposing the architecture of an educational game designed for middle and high school students. The game will be introduced and tested in classes to support educational projects for inclusion through art history education.

### **PAROLE CHIAVE**

Serious Game; Fine Arts; Media and Visual Culture; Inclusion and Equality; Virtual Assessment

## **1. INTRODUCTION**

In recent years, themes such as *inclusion* and *equality* have occupied a crucial position in the international agenda, even with the rapid digitization of society. While the development of digital educational approaches, the availability of new resources, and the growth of European initiatives to face social injustice have increased in interest among government institutions and the private sector, these efforts still need to be challenged in daily practices. Indeed, virtual and real violence against women and other marginalized groups and embedded stereotypes in Western culture continues to grow<sup>1</sup>. To raise awareness and shift societal norms is to provide adequate educational resources to young individuals still receptive to learning and establishing moral values [1][2].

Increasingly, leading museums and cultural institutions around the globe adopt digital tools to support learning for young visitors, but how can we assure the same educational strategies in places where access to seats of culture is not guaranteed? In order to ensure equal access to resources that educators can integrate both traditional knowledge and the latest ethical matters, in this work we propose the game design of a Serious Game dedicated to the study of stereotypical representations of minorities’ identity through critical education in visual culture and fine arts.

Traditionally, issues related to stereotypical roles have been widely discussed in Critical Cultural Studies and Postcolonial Studies applied to Fine Arts, and the review of the literature available in these fields exceeds the scope of this work[14][13]. However, at least two main approaches to education on inclusion in the arts can be traced from the reference literature. These approaches are discussed in more detail in section 2, together with a brief state of the art of knowledge background involved in the development of the game design.

In section 3, we show how we apply the theoretical background to the design of the game prototype, exploring its architecture, expected learning outcomes, the serious game approach involved, and exemplary case studies. In section 4, we briefly conclude by indicating the next stages of development of the project and the upcoming experimental setting, in

---

<sup>1</sup> <https://eige.europa.eu/gender-statistics/dgs>



order to assess the game experiences of adolescents and young adults, as well as the educators in the school context. Through our design, we intend to make the knowledge acquired useful in everyday life, characterized by the dominance of visuality.

## 2. ENGAGED EDUCATION, VISUAL CULTURE, AND VIDEO GAMES

Scholars in the Social History of Art and Cultural Studies have heightened the role of Cultural Heritage and visual experiences as a form of self-awareness. In this perspective, in the cultural system, artifacts are entangled in relationships with other objects and subjects, places, processes, and practices. Past events depicted in a work of art acquire relevance by understanding the context of their origin and their genealogy, including the history of how concepts such as gender, race, and class have been conceived over time. Given the possibility offered by open data and query systems, digital tools are configured as natural means to support this inquiry in education practices [9][10].

To strongly emphasize the role of education in critical reading of artistic and intellectual productions, among many others, it is important to cite the American philosopher bell hooks. From her situated experience as a black woman, the scholar starts a general reflection on the importance of practices of cultural re-visioning, starting from a marginal, subaltern point of view. For hooks, teaching is a performative act of liberation that, far from reinforcing systems of exclusion, must offer the world a concrete space for change. Thus, classrooms become laboratories of virtuality in which the concepts of theory and practice cannot be addressed in isolation [11]. Therefore, crucial for the development of this project is the concept of *engaged pedagogy*: the conviction that teaching practice is not only about the sharing of information and notions, “but to share in the intellectual and spiritual growth of our students” ([12]:14). Bell Hooks’ “standpoint” approach emphasizes the importance of teaching critical awareness, particularly for those marginalized and discriminated individuals who suffer from the conditions of intellectual isolation typical of the peripheries of the world. Furthermore, this approach questions the power dynamic present in the teacher-student relationship and thus seeks to construct a learning environment where both sides can learn from each other. This is important because hierarchical constraints can occur not only between teachers and students but also between students from different social classes and from different nationalities and cultural backgrounds. Thus, the bell hook approach restores the idea that education plays an essential role in the transformation of society, and educators should be selective in determining students’ experiences. A proper selection of content is needed to enhance individual and social growth and lead to the empowerment of social minorities in everyday circumstances.

Following her path, the approach we intend to propose in this study focuses on the visual arts. In the tradition of Feminist Art History, the first level of analysis dealt with the under-representation of female artists and ethnic minorities within the Western artistic canon [13]. Within the same framework, Freed observes that “women artists have begun to be written into art history, but often appear as if injected into a story about *great men*” ([13]:158). Exemplary of this attempt is *Die Kunst der Gesellschaft 1900-1945*, a recent curatorial project of the Neue Nationalgalerie zu Berlin<sup>2</sup>, which tries to overcome the glass roof of Modern and Contemporary Art by showing at least a female exponent for each dowel of the thematic path ranging from 1900 to 1945. Therefore, in this context, educating means asking if this is enough and bridging the gap constituted by access to these same sources so that the museums of large European capitals are not a place of privileged access to knowledge. Some works that will be investigated within our game are inspired by this historical journey.

The second concern focuses on the stereotypical representation of genders and ethnicities in artistic representations over time. This theme is extensively investigated in the literature, and an exhaustive review of this topic is beyond the scope of this work. However, it should be noted that this educational approach, which places the reading and interpretation process at its core, works as a support not only for the institutional learning scopes, that is, understanding the History of Art, but also supports the idea that knowledge is an integral part of our daily social life. We live in a world where visual representation (drawings, pictures, photos) and videos spreading in social networks such as Instagram and TikTok play an even larger role in shaping morality in young individuals<sup>3</sup>. Thus, educating students to read the implicit messages hidden behind a famous picture stimulates a critical reflection both about the past and about the immediate present [14].

The same consideration led us to the third level of inquiry: how to educate young scholars to challenge the stereotyped representations of target groups within the mainstream media, the same media that reinforces and encourages the permanence of such stereotyped characterizations? [20]. This is particularly burdensome when looking at video game products. Just to name a few, the game *Custer’s Revenge* (Mystique 1982) was one of the first examples of video games with stereotypical and racist representations within a video game. Despite the graphical limitations dictated by technical constraints over time, the game published for the Atari 2600 places the player in the role of George Armstrong Custer, an

---

<sup>2</sup> Visit the web site: <https://www.smb.museum/museen-einrichtungen/neue-nationalgalerie/ausstellungen/detail/die-kunst-der-gesellschaft> (last visit 28/02/2023).

<sup>3</sup> <http://manovich.net/index.php/projects/instagram-and-contemporary-image> (last visit 28/02/2023)

officer in the United States Army during the American Civil War, dodging naked the arrows shot by Native Americans to reach a girl tied up in order to rape her. Games such as *Night Trap* (Digital Pictures 1992), *The Grand Theft Auto* series, or *Mortal Kombat* (Midway Games 1992), which show extremely violent content, are also extremely popular. How can it be possible to hack the connection between violence and digital games and appropriate a medium that arouses interest due to its hurtful elements to educational favor?

Several recent studies in Gamification show that the educational power of digital games can be fruitfully involved in a process of raising awareness on issues of equity and injustice [21]. Gamification, or the use of game elements and game design techniques in a non-game context [5], has been shown to have positive consequences in raising the awareness of the user while also making them have fun. The inclusion of playful elements in daily or work routines engages users in their tasks and puts them in a state of flow [6]. As a result, the integration of playfulness into the learning process is becoming crucial for digital natives. Thanks to gaming elements, the task is perceived as pleasing and easy to remember, and the introduction of game quests has shown to be an effective tool for improving the acquisition of specific skills [7]. Understanding game mechanics can also be considered a form of education, and it was only a matter of time before video games became a part of institutionalized education [8]. Despite the tendency to sensationalize the issue of violence in video games more often than the effectiveness of educational methods related to them, we can state that the negative effects of video games would be more sporadic than commonly believed. In fact, studies show that there are more benefits than disadvantages in their consumption [22]. In our game, we question the privileged viewpoint of young gamers, challenging them to question their beliefs about the world and our past.

### **3. BREAKING THE CANVAS: GAME DESIGN AND MAIN PURPOSES**

*Breaking the Canvas* is a quiz game designed to gain awareness among young scholars of stereotypes across Visual Culture and the History of Fine Art. To this aim, the game shows several famous paintings and challenges players to answer nontrivial questions related to the artwork. The purpose is twofold: on the one hand, to test the familiarity with the visual artifact by the player at different levels and on the other hand, to train the critical evaluation of the same object by activating a process of deconstruction of the image represented as a whole. The young scholar must notice the details hidden in the image and provide a single valid answer among the options. The difficulty lies above all in the fact that some of the options provided can be considered correct as well, but superficial in their critical readings of the artwork.

The first level of the game is dedicated to the recognition of the main attributes of the picture, such as data relating to the creation of the piece of art (year, historical framework, and artistic movement), information about the author, and the current location of the works. In this first phase, deep observation and critical reflection are not required by the students, as it is assumed that information is gained from the textbooks and/or made available by the teacher in advance. Descriptive documentation and metadata associated with the works currently cataloged within the platform are also made available to students as they progress through the game. As a bonus for correct answers, more information appears on the screen as a new piece of the puzzle immediately after the answer is provided, and in association with some salient elements depicted in the scene. This is not accidental, and the student should pay attention to the clues as they will become crucial in the second phase of the game. This ensures that the students answer the questions and focus on the details disclosed in the image, memorizing the contents continuously in order to pass levels and access new ranks.

The second level of the game rouses the recognition of more evident gender and ethnicity storytelling elements in the works of art. This level relies on the players' ability to store, understand and select information about the clues within the scenes, whether they are related to gender differences, race, or sexual orientation. Here the student is guided through thematic pathways called *frames* in which works of art from different historical periods and artistic currents are associated and compared with each other. This method makes the acquisition of theoretical skills in image reading immediately applicable, training the student to find points of continuity and break between the representations (see Figure 1). Just as with questions at the first level, after the correct answer, players will have excess textual elucidation about the history behind the image they are facing and the associated frame. The goal is to stimulate curiosity, reflection, and critical thinking around themes and works of art.

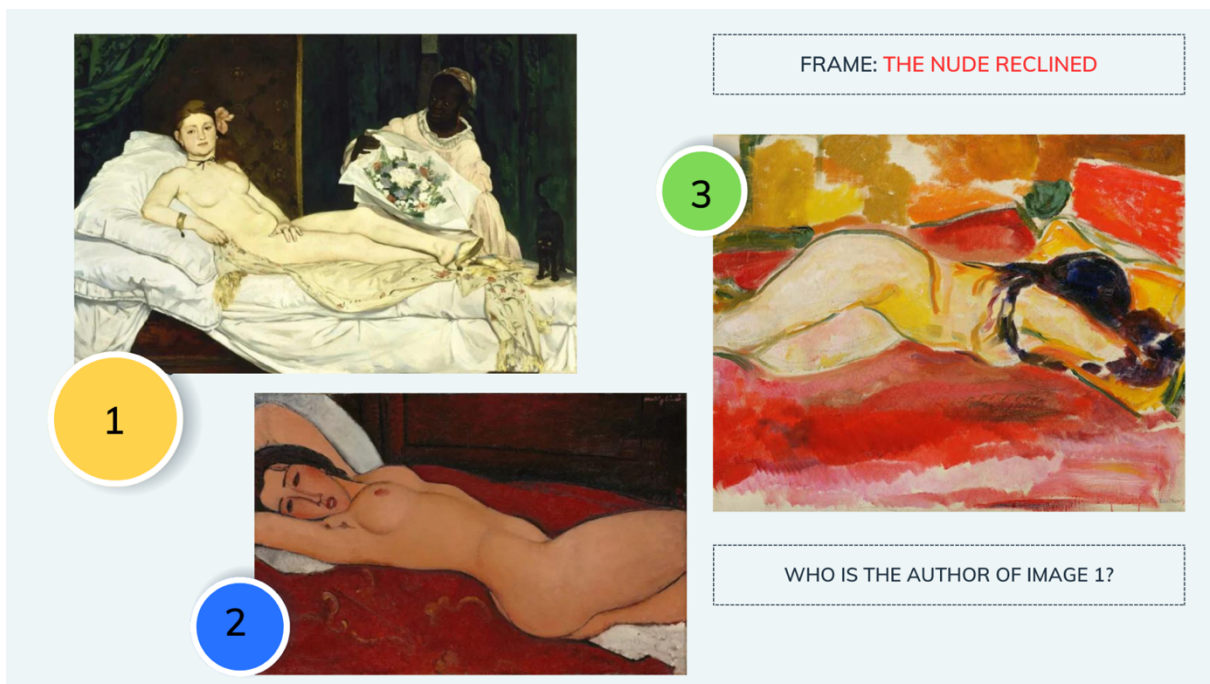


Figure 1. The frame "The nude reclined" shows three famous paints depicting figures of women in a very traditional pose. She is a goddess, odalisque, or a beautiful woman abandoned on a surface. From the left to the right: Eduard Manet, Olympia, 1856, Musée d'Orsay, Paris; Amedeo Modigliani, Reclining Nude, 1917, from a private collection; Edvard Munch, Reclining Nude, 1912, Hamburger Kunsthalle, Hamburg.

Thanks to the framing storytelling, the students become familiar with the authors and the paintings as well as with the subjects and themes that have been established as canonical in Western figurative art. In the case of *the nude reclined* frame, for instance, the student acquires the ability to grasp the differences and similarities and the evolution of how the same subject is depicted in different historical periods, or by contemporaries influenced by different artistic waves (as in the case of Munch and Modigliani).

The approach we used mirrors the exploitative nature of the *explorer* archetype described in Bartle's taxonomy [22], constantly on the hunt for new discoveries and information. In order to provide support during the experience both for students and teachers in the classroom, the gamer is accompanied by a travel companion throughout the levels (a robot-shape figure named Helmo who will guide the gamer through the platform affordances). The so-called *call to adventure*[18] is Helmo's request to learn how to interact with others without generating stereotypes of any sort. The presence of the robot is useful in creating a relationship within the game world, orientating the experience to an actual scope, and increasing the player's motivation to pursue the narrative progress that will emerge from their relationship. Every time the player answers correctly, Helmo will show joyful feedback and a message of congratulations, while if the player answers incorrectly, the robot will display a sad emoji and a message that will try to motivate the player to improve their responses. The choice to use emoji is driven by the desire to use a universal medium that can easily and immediately show the expressiveness experienced by our travel companion, making it clear to the user whether what they see on the screen is a positive or negative event. The player serves as a mentor on Earth for the buddy, but as the story progresses, the player also learns the concepts presented by the game alongside the robot.

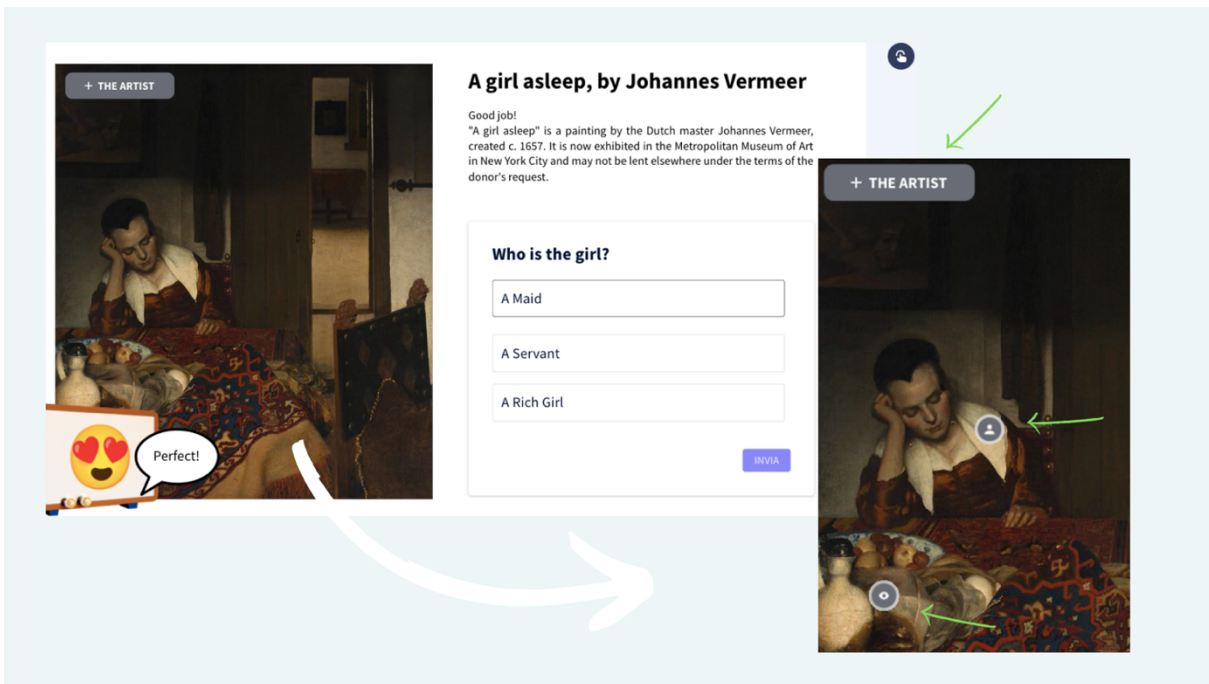


Figure 2. An example of a frame that questions the figurative trope of the maid, as depicted in a famous painting. On the right side of the image, indicated by the green arrow, the suggestions that can appear directly on the image to provide the students with some insights into the artist and the work.

Upon completion of the game, students will be tested on their mastery of stereotype analysis through an Extra Mode featuring AI-generated paintings. The aim here is to help both players and educators in testing the improvements about the topics. The frames presented in this section will be generated using *Dall-E 2*, an AI algorithm capable of creating images based on textual descriptions. *OpenAI*, as the creator of the algorithm, is dedicated to improving the inclusivity of its product by minimizing the risk of creating stereotypical biases when generating images [24]. However, the possibility of encountering cultural biases still present in the relatively young *Dall-E 2* algorithm is quite frequent in processing generic information [25]. The system can effectively comprehend human language and accurately produce user requests in various styles. The algorithm tends to make the most errors in generating images of complex scenes, where it struggles to accurately position all the described elements, resulting in incomplete images or misplaced elements with incorrect relationships [26]. To generate an effective educational image, the textual input must contain only a few precisely described elements. This not only prevents the algorithm from becoming confused by handling too many elements but also allows us to focus the students on the most important lessons within each proposed image. This tool was chosen to generate images to overcome two possible complications: 1. To provide contemporary examples without risk in violating the copyright associated with the reproduction of advertising and media materials; 2. To show how human biases can have a direct impact on the production of new knowledge and technologies.

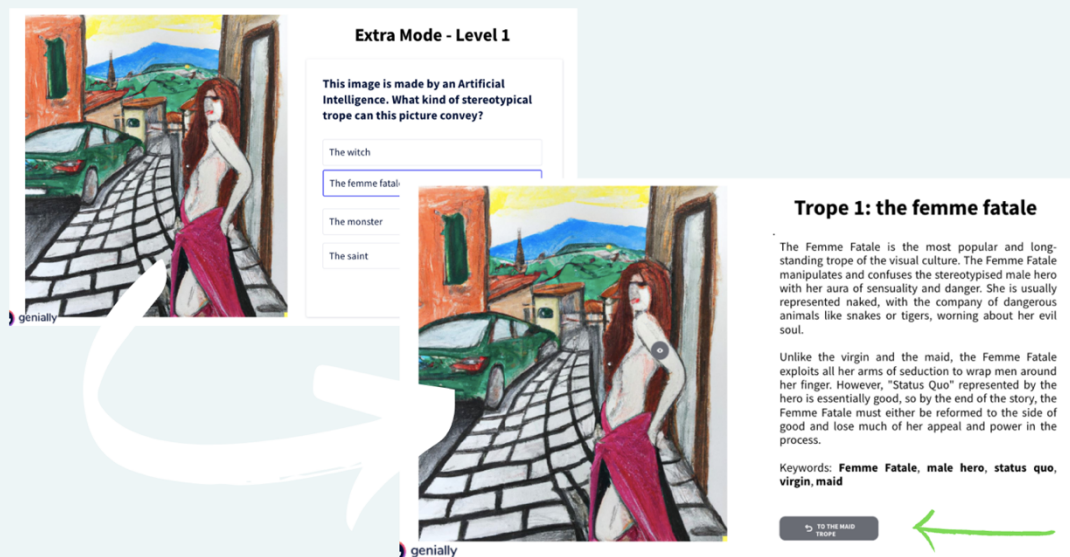


Figure 3. An example of image generated by DALL E 2 to convey the trope of the femme fatale. It is possible for the students to navigate from the present page back to past pages by interlinks, as indicated by the green arrow.

#### 4. CONCLUSION AND FUTURE WORK

The project seamlessly integrates into art history courses, supplementing traditional academic teachings with a heightened focus on education and sensitivity. The game demo is currently available in the Genially environment, a web-based tool that provides templates for Gamification initiatives, easy to use both for teachers and scholars. Genially is user-friendly and highly customizable, granting developers the freedom to create even new educational content that supports hybrid teaching methods [23]. We employ this platform after a thorough examination of the target audience and the technological infrastructure provided by academic laboratory settings. It is common for such facilities to possess outdated equipment or lack the capabilities to support ultra-detailed and complex gaming experiences. Moreover, Genially is accessible on any device that can connect to an internet network, offering universal access at any time and place.

As further improvement of the present study, we plan to make the game independent from the Genially platform by developing an interactive environment designed to support the implementation of a Knowledge Base to expand the pool of pictures with sources from real public database, such as that currently provided by the Prado Museum in Madrid. The use of resources from museums and institutions is significant, because it underlines the effort to make cultural resources accessible inaccessible for many students all around the globe. Finally, we plan to test the demo first on high school students from different study curricula between Sicily and Liguria.

#### REFERENCES

- [1] E. Chung and E. Turiel, Adolescents' judgments about resource inequality involving group disparities, *Journal of Experimental Child Psychology* 218, (2002).
- [2] S. J. Brose, C. Conry-Murray, E. Turiel, Adolescents' reasoning about parental gender roles, *The Journal of Genetic Psychology* 174 (2013) 207–224.
- [3] D. Michael, S. Chen, *Serious Games: Games that Educate, Train and Inform*, Thomson Course Technology, 2006.
- [4] C. C. Abt, *Serious Games*, New York: Viking Press, 1970.
- [5] K. Werbach, D. Hunter, *For the Win: How Game Thinking Can Revolutionize Your Business*, Wharton Digital Press, 2012
- [6] M. Csikszentmihalyi, *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, Springer Netherlands, 2014.
- [7] S. Iacono, M. Coccoli, D. Zolezzi, G. Vercelli, Gamification di un video-corso sulle competenze informatiche di base: da Gli Irriunciabili a Star Words, in: *Teach Different! Proceedings to EMEMITALIA2015*, Genova University Press, 2015, pp. 275–278.
- [8] T. Chatfield, *Fun Inc: Why Games are the Twentyfirst Century's Most Serious Business*, Virgin, 2011
- [9] S. Anastasi, A. De Domenico, M. Nicolosi-Asmundo, (2022) Egith-engendering data. Steps towards an ontology for the representation of gender in cultural heritage, in: *1st Workshop on Artificial Intelligence for Cultural Heritage, AI4CH 2022*, CEUR-WS.org, 2022, pp. 13–25



- [10] P. Jaskot B., Digital art history as the social history of art: Towards the disciplinary relevance of digital methods, 2019.
- [11] b. hooks, *Insegnare a trasgredire, L'educazione come pratica della libertà*. Meltemi (2020).
- [12] b. hooks, *Teaching to transgress*. Routledge (2014).
- [13] L. Nochlin, Why have there been no great women artists?, *ARTnews* (1971).
- [14] K. Freedman, Interpreting gender and visual culture in art classrooms, *Studies in Art Education* 35 (1994) 157–170.
- [15] C. McCormack, *Women in the Picture: Women, Art and the Power of Looking*, Icon Books, 2021.
- [16] M. Prensky, *Digital Game-based Learning*, McGrawHill, 2001.
- [17] A. Kapriyelov, Representation of women in video games, *Liverpool John Moores University/Liverpool: Liverpool Screen School* 37 (2015).
- [18] L. Castillo-Cuesta, Using Genially Games for Enhancing EFL Reading and Writing Skills in Online Education, *International Journal of Learning, Teaching and Educational Research* 21 (2022).
- [19] C. Vogler, *The Writer's Journey: Mythic Structures for Screenwriters and Storytellers*, M. Wiese Productions, Studio City, CA, 1992
- [20] D. Williams, N. Martins, M. Consalvo, J. D. Ivory, The virtual census: Representations of gender, race and age in video games, *New media & society* 11 (2009) 815–834.
- [21] T. G. Deskins, Stereotypes in video games and how they perpetuate prejudice, *McNair Scholars Research Journal* 6 (2013) 5.
- [22] M. Rickards, F. Vannucchi, *Il videogioco. Mercato, giochi e giocatori, Libri per la comunicazione*, Mondadori Università, 2013
- [23] R. Bartle, *Hearts, clubs, diamonds, spades: Players who suit MUDs* (1996)
- [24] N. Hermita, Z. H. Putra, J. A. Alim, T. T. Wijaya, S. Anggoro, and D. Diniya, 'Elementary Teachers' Perceptions on Genially Learning Media Using Item Response Theory (IRT)', *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, vol. 4, no. 1, pp. 1–20, 2021.
- [25] OpenAI, 'Reducing bias and improving safety in DALL·E 2'. 2022. [Online]. Available: <https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2>
- [26] F. Offert and T. Phan, *A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space*, 2022
- [27] G. Marcus, E. Davis, and S. Aaronson, *A very preliminary analysis of DALL-E 2*, 2022.

# CoMOntology. Towards An Ontology for the Chronology of Mediterranean archaeologies: a model for the digital memory

Rodolfo Brancato<sup>1</sup>, Marianna Figuera<sup>2</sup>, Marianna Nicolosi-Asmundo<sup>3</sup>,  
Daniele Francesco Santamaria<sup>4</sup>, Paola Venuti<sup>5</sup>, Giuseppe Zappalà<sup>6</sup>

<sup>1</sup> Dipartimento di Studi Umanistiche, Università di Napoli Federico II, Italia – rodolfo.brancato@unina.it

<sup>2</sup> Dipartimento di Scienze Umanistiche, Università degli Studi di Catania, Italia – marianna.figuera@unict.it

<sup>3</sup> Dipartimento di Matematica e Scienze Computazionali, Università degli Studi di Catania, Italia – marianna.nicolosiasmundo@unict.it

<sup>4</sup> Dipartimento di Matematica e Scienze Computazionali, Università degli Studi di Catania, Italia – daniele.santamaria@unict.it

<sup>5</sup> Dipartimento di Scienze Umanistiche, Università degli Studi di Catania, Italia – venuti.paola@gmail.com

<sup>6</sup> Dipartimento di Scienze Umanistiche, Università degli Studi di Catania, Italia – giuseppe.zappala@unict.it

## ABSTRACT\*

Il presente lavoro – sviluppato nell’ambito delle attività di ricerca sulle *digital humanities* presso l’Università di Catania – è dedicato alla progettazione e allo sviluppo di un’ontologia per il Web Semantico con l’obiettivo di modellare la relazione spazio-tempo desumibile dai dati *legacy* di interesse archeologico. Tale ricerca, ancora in corso, è attualmente incentrata sulla datazione relativa e assoluta delle testimonianze archeologiche delle civiltà del Mediterraneo orientale e centrale nel corso della Preistoria, dal X alla fine del I millennio a.C., come rappresentato dalle tavole cronologiche di C. Broodbank. L’approccio proposto, oltre alla periodizzazione assoluta e relativa, mira a delineare alcune altre questioni come le fluttuazioni climatiche.

## PAROLE CHIAVE

Cronologia; archeologia; CIDOC-CRM; web semantico; legacy data

## 1. INTRODUZIONE E STATO DELL’ARTE

L’integrazione digitale tra dati archeologici, *legacy data* e le nuove ricerche deve fare i conti con gli elementi archetipici della realtà, ovvero spazio e tempo [2]. L’avvento e l’applicazione di tecniche come la datazione al 14C o la dendrocronologia non hanno ridotto in modo significativo la necessità di cronologie basate sulla cultura materiale, così come asserito anche in [8], in cui cultura materiale e cronologia appaiono indissolubilmente legate da una prospettiva ontologica. Oggi, infatti, risulta ancora evidente come i diversi metodi di datazione non portino direttamente a un quadro cronologico coerente per il contesto territoriale di riferimento; di conseguenza, uno dei problemi principali da affrontare resta ancora la sincronizzazione dei dati cronologici provenienti da fonti diverse, seguendo gli sviluppi della ricerca sostenibile in ambito europeo [5].

Il nostro contributo mira alla progettazione e allo sviluppo di un’ontologia per il Web Semantico per la modellazione delle relazioni tra i dati *legacy* e i relativi aspetti cronologici desunti dalle testimonianze archeologiche delle civiltà del Mediterraneo. Il problema delle rappresentazioni ontologiche che trattano il concetto di tempo è già stato affrontato sia dal *World Wide Web Consortium (W3C)*<sup>1</sup> che da CIDOC *Conceptual Reference Model (CRM)*<sup>2</sup>, fino ad un’applicazione del metodo *fuzzy* per preservare il concetto di flessibilità cronologica [6]: tra le più recenti applicazioni pratiche di tale metodo vi è, infatti, quella legata all’influenza che il rapporto spazio-tempo ha sulla definizione di cultura materiale [8]. A questo proposito anche il CIDOC-CRM, versione 7.2.2, propone primitive relazioni temporali basate su confini *fuzzy* che consentono l’accurata interposizione di entità temporali, considerando la sfocatura intrinseca dei confini temporali. Il modello proposto da CIDOC-CRM, tuttavia, non tiene in considerazione l’aspetto cronologico in termini archeologici, in particolare nel contesto mediterraneo, prestandosi così poco ad applicazioni in tal senso.

---

\* Il lavoro è frutto del confronto tra gli autori nelle fasi di ideazione, progettazione, stesura e revisione. In particolare si devono ad 1 Introduzione e stato dell’arte: Rodolfo Brancato e Marianna Figuera, 2 Un modello per la memoria digitale del Mediterraneo: Marianna Nicolosi Asmundo, Daniele Francesco Santamaria, Paola Venuti, Giuseppe Zappalà, 3 Conclusioni e sviluppi futuri: Rodolfo Brancato, Marianna Nicolosi Asmundo

<sup>1</sup> World Wide Web Consortium (W3C), <https://www.w3.org/>

<sup>2</sup> CIDOC Conceptual Reference Model (CRM), <https://www.cidoc-crm.org/>



Per tale ragione, nel desiderio di costruire un modello uniforme, si è rivista la cronologia del Mediterraneo usando un approccio ontologico, partendo da precedenti lavori sulle periodizzazioni relative (PeriodO<sup>3</sup>) [10] e collegandoli a un ambito più ampio di fluttuazioni climatiche e periodizzazioni assolute.

## 2. UN MODELLO ONTOLOGICO PER LA MEMORIA DIGITALE DEL MEDITERRANEO

Il modello ontologico proposto è stato definito a partire dalla tavola cronologica di C. Broodbank [3], dalla quale sono state individuate le seguenti classi:

- **Absolute Chronology**, che comprende le ere geologiche e gli eventi climatici che si sono verificati nel corso di esse. Ci siamo concentrati sul periodo del Quaternario, suddiviso nelle due epoche del Pleistocene (2,58 milioni di anni fa - 11700 anni fa) e dell'Olocene (11700 anni fa - in corso). All'interno delle fluttuazioni climatiche, è possibile distinguere le varie fasi secondo la classificazione di Blytt-Sernander [7], in modo da poter definire fasi cronologiche assolute all'interno delle quali rientrano le varie cronologie relative delle diverse aree.
- **Relative Chronology**, suddivisa in Età della Pietra, Età dei Metalli ed Età Classica con relativi sottoperiodi. Sono stati identificati due *datatypes* 'start date' e 'end date' per la classificazione automatica attraverso il *reasoner*.
- **Area**, suddivisa in sottoclassi specifiche per ogni area geografica del Mediterraneo, all'interno delle quali è possibile riconoscere le singole zone geografiche, periodizzate secondo la relativa cronologia.
- **Archaeological evidence**, che raccoglie testimonianze archeologiche, come manufatti e qualsiasi tipo di prova materiale che ci è stata tramandata grazie a particolari eventi.
- **Reference**, è legata alla proprietà che indica i riferimenti bibliografici, con una sotto-proprietà specifica per i riferimenti Wiki.

Il modello E-R così costruito (Figura 1) ha così permesso di definire le relazioni fra le classi e la conseguente determinazione di rapporti fra gli individui esplicitati, ponendo in relazione gli individui della cronologia assoluta con quelli della cronologia relativa per mezzo delle periodizzazioni climatiche verificatesi, e giungendo alla relazione fra cronologia relativa, evidenze archeologiche e aree geografiche. L'ontologia è stata progettata facendo uso dell'editor *Protégé*<sup>4</sup>.

---

<sup>3</sup> PeriodO. A gazetteer of periods for linking and visualizing data, <https://perio.do/en/>

<sup>4</sup> <https://protege.stanford.edu/>

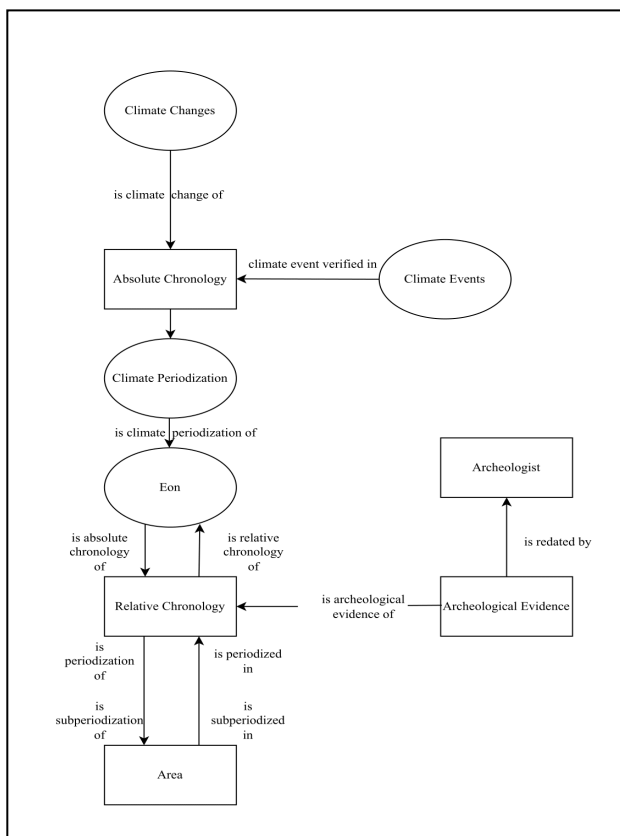


Figura 24. Modello Entità-Relazione (E-R) del progetto CoMOntology.

Le periodizzazioni e le sotto-periodizzazioni di ciascuna area sono state trattate come individui specifici, indicando a) la data di inizio e la data di fine attraverso la definizione di opportune relazioni e b) la fase cronologica di appartenenza attraverso la proprietà di cronologia assoluta che esprime l'individuo della fase cronologica specifica. Per superare il problema della sovrapposizione cronologica, sono state create apposite regole *Semantic Web Rule Language (SWRL)* per gestire le diverse fasi cronologiche all'interno dello stesso periodo. Queste regole permettono di calcolare periodi e sottoperiodi delle diverse aree in modo da determinare – all'interno delle fluttuazioni climatiche – la sovrapposizione prodotta dalle categorizzazioni archeologiche, e quindi di definire – all'interno della cronologia assoluta – una mappa di luoghi ed eventi che vada oltre la facies archeologica e che tenga conto dei vari processi evolutivi avvenuti nelle singole aree (Figura 2).

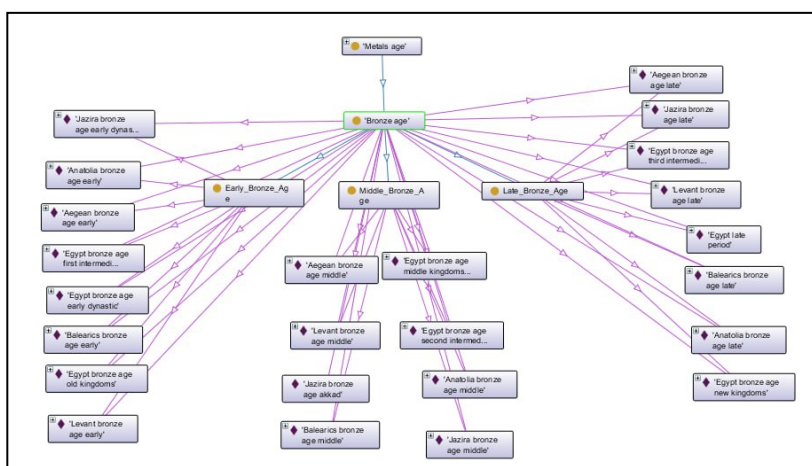


Figura 25. Individui della classe *Bronze age*.

Le diverse classi e sottoclassi individuate sono state messe in relazione attraverso specifiche *object-properties*, attraverso cui si è in grado di definire relazioni fra gli individui, permettendo al ragionatore automatico di inserirle, mostrando una corrispondenza unilaterale o bilaterale fra gli stessi. Di particolare rilievo per gli obiettivi del progetto è l'*object-properties*

che collega gli individui appartenenti alle cronologie relative con le fluttuazioni climatiche verificatesi nello specifico arco cronologico. (Figura 3)

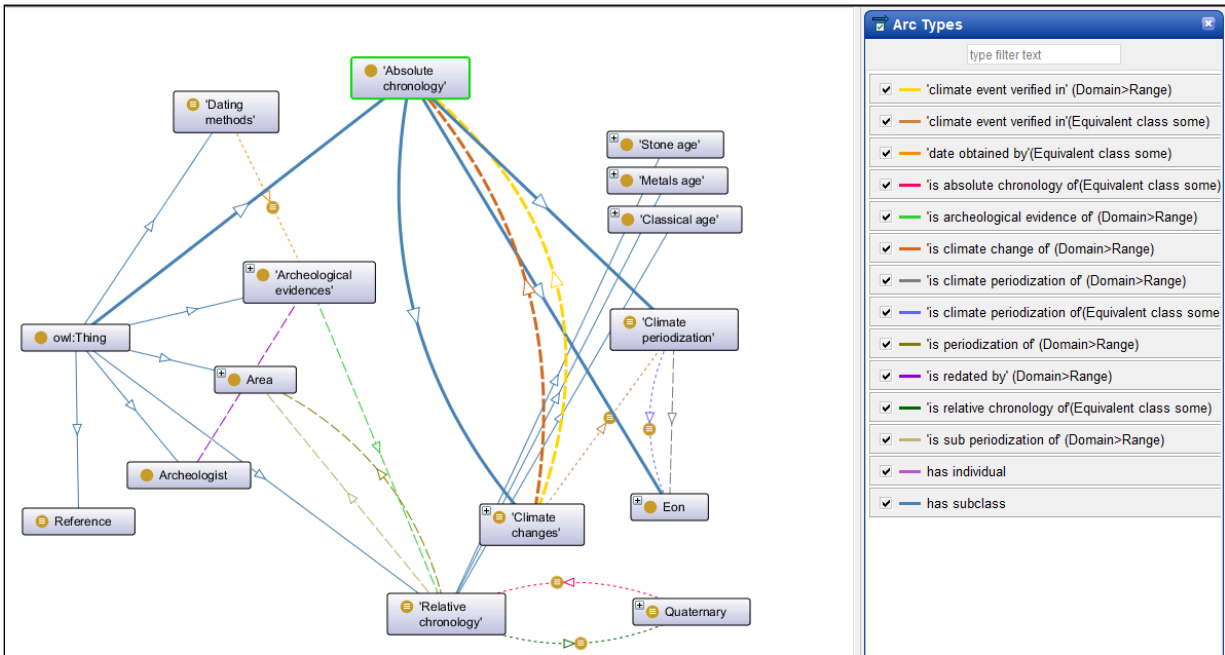


Figura 26. Visualizzazione ad arco delle *object-properties*.

Inoltre, sono state introdotte alcune restrizioni per dedurre coerentemente le informazioni cronologiche attraverso il ragionatore automatico *Pellet*, che è stato in grado di rispondere correttamente alle asserzioni attese, in particolare nel definire relazioni fra gli individui delle diverse classi collegati da specifiche *object properties*.

Come si evince dalle regole imposte (Figura 4), è possibile collegare differenti periodizzazioni relative di specifiche aree con l'era geologica di appartenenza, così da rappresentare una mappa dei diversi processi evolutivi che hanno interessato le singole aree nel medesimo arco cronologico di interesse, rispettando il sistema climatico proposto da Blytt-Sernander.

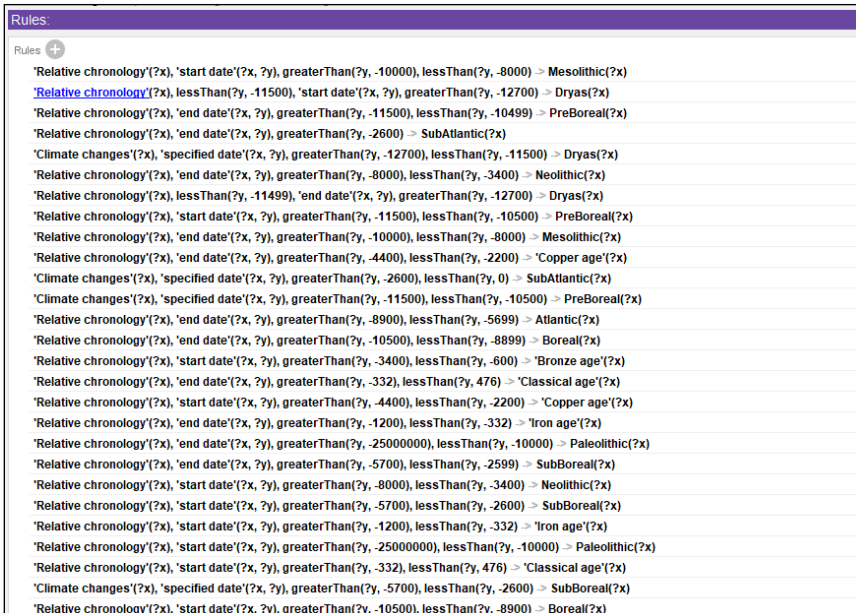


Figura 27. Regole *SWRL* per gestire le diverse fasi cronologiche all'interno dello stesso periodo.

### 3. CONCLUSIONI E SVILUPPI FUTURI

Questo lavoro si occupa della progettazione e allo sviluppo di un'ontologia web che rappresenti le relazioni tra dati *legacy* e i relativi aspetti cronologici. La realizzazione di un modello ontologico è attualmente in fase di sviluppo, con numerosi aspetti ancora da considerare. Il modello, infatti, dovrà essere integrato con le ontologie attualmente disponibili, quali

CIDOC CRM, e le relative estensioni, quale ad esempio iDAI.Chronontology<sup>5</sup>, disponibili per le cronologie dell'Europa centrale. L'ontologia costruita su tale modello rappresenterà la prima fase verso la realizzazione di una piattaforma web per l'interrogazione e la visualizzazione dei dati raccolti e l'integrazione con altri LOD, a partire dai testi condotti su Sicilia [1] e Malta [10].

## BIBLIOGRAFIA

- [1] Brancato, Rodolfo. «How to access ancient landscapes? Field survey and legacy data integration for research on Greek and Roman settlement patterns in Eastern Sicily». *Groma* 4 (2019): 75-106.
- [2] Brancato, Rodolfo, Nicolosi-Asmundo, Marianna, Pagano, Grazia, Santamaria, Daniele Francesco, Uccchino, Salvatore. «Towards an ontology for investigating on archaeological Sicilian landscapes». In *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage*, C. Meghini, A. Poggi (eds.), Vol. 2375, pp. 85-90. Rome: ODOCH 2019, 3 June 2019, CEUR Workshop Proceedings.
- [3] Broodbank, Cyprian. *The Making of the Middle Sea: A History of the Mediterranean from the Beginning to the Emergence of the Classical World*. Oxford: Oxford University Press, 2013.
- [4] Cantone, Domenico, Nicolosi-Asmundo Marianna, Santamaria Daniele Francesco, Trapani, Francesca. «OntoCeramic: an OWL ontology for ceramics classification». In *Proceedings of the 30th Italian Conference on Computational Logic*, D. Ancona, M. Maratea, V. Mascardi (eds.), Vol. 1459, pp. 122-127. Genova: CILC 2015, July 1-3 2015, CEUR Workshop Proceedings.
- [5] Cichocki, Otto, Bichler, Max, Firneis, Gertrude, Kutschera, Walter, Müller, Wolfgang, Stadler, Peter. «The Synchronization of Civilizations in the Eastern Mediterranean in the Second Millennium BC: Natural Science Dating Attempts». In *Tools for Constructing Chronologies. Lecture Notes in Statistics*, C.E. Buck, A.R. Millard (eds), 177, pp. 83-110. London: Springer, 2004, <https://cordis.europa.eu/project/id/HPCF-CT-2000-00075-02/it>
- [6] Figuera, Marianna. «A fuzzy approach to evaluate the attributions reliability in the archeological sources». *International Journal on Digital Libraries* 22(3). (London:Springer, 2021): 289-296.
- [7] Foster, Richard. *Glacial and Quaternary Geology*. New York: John Wiley, 1971.
- [8] Niccolucci, Franco, Hermon, Sorin. «Formally defining the time-space-archaeological culture relation: problems and prospects». *Archeologia e Calcolatori* 28. (2017): 93-108.
- [9] Orton, Clive, Tyers, Paul, Vince, Alan. *Pottery in Archaeology, Cambridge Manuals in Archaeology*. Cambridge: Cambridge University Press, 1993.
- [10] Parkinson, Eóin W., McLaughlin, Rowan T., Esposito, Carmen, Stoddart, Simon, Malone, Caroline. «Radiocarbon Dated Trends and Central Mediterranean Prehistory». *Journal of World Prehistory* 34 (2021): 317–379, <https://link.springer.com/article/10.1007/s10963-021-09158-4#citeas>
- [11] Rabinowitz, Adam. «It's about time: historical periodization and Linked Ancient Word Data». *ISAW Papers* 7: 22, <http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/rabinowitz/>

---

<sup>5</sup> iDAI.Chronontology, <https://chronontology.dainst.org/>

# Dematerializzazione e sostenibilità: prospettive ideali e problemi concreti

Flavia Bruni<sup>1</sup>

<sup>1</sup> Università Gabriele D'Annunzio, Chieti-Pescara, Italia – [flavia.bruni@unich.it](mailto:flavia.bruni@unich.it)

## PAROLE CHIAVE

Dematerializzazione; conservazione del digitale; ecosostenibilità; ambiente

## 1. OBIETTIVE PREMESSE

All'inizio del terzo millennio, le questioni irrisolte in territorio digitale restano molte. La cronica mancanza di coordinamento generale, particolarmente evidente nel campo delle digital humanities, ha spesso segnato il destino di progetti arenatisi su ostacoli già fatali ad analoghe imprese precedenti; un fattore che appare oggi determinante nel minare la credibilità dei progetti digitali è riconducibile al problema della sostenibilità su molteplici piani interconnessi: dal punto di vista delle decisioni strategiche delle istituzioni ospitanti; dal punto di vista meramente tecnologico della disponibilità delle piattaforme e infrastrutture; infine, con un'urgenza che esula dal mero tecnico, dal punto di vista ambientale. Seppure piuttosto recente nell'ambito del dibattito su queste tematiche, l'interrogativo riguardo l'impatto delle tecnologie informatiche sull'ambiente e sul clima non ammette più deroghe visto il rischio concreto di pregiudicare la stessa esistenza del pianeta. La crisi ambientale in atto si impone ormai costantemente e in varie forme alla nostra attenzione, esigendo una radicale revisione della nostra organizzazione quotidiana in quanto singoli e in quanto società. Un'accresciuta sensibilità nei confronti delle tematiche ambientali è permeata fino all'interno della Carta costituzionale, i cui articoli 9 e 41 sono stati modificati l'8 febbraio 2022 introducendo tra i principi fondamentali la tutela dell'ambiente, della biodiversità e degli animali<sup>1</sup>. All'interno della sconfinata letteratura sull'impatto rivoluzionario determinato dall'avvento del digitale su molteplici piani della nostra vita, l'aspetto specifico dell'incidenza delle tecnologie digitali sull'emergenza energetica e climatica pare essere rimasto un terreno finora poco battuto. Partendo dall'analisi semantica della documentazione relativa alla cosiddetta dematerializzazione in atto nella Pubblica Amministrazione, questo contributo propone alcuni spunti per una riflessione costruttiva a fronte di criticità cui è stata finora riservata una scarsa attenzione.

## 2. L'IDEA DIGITALE

Nell'ambito di ciò che si definisce conservazione del digitale (digital preservation) si distinguono più filoni, in base alle sfaccettature che la disciplina assume in relazione ai diversi aspetti interessati, mentre il concetto stesso di conservazione non è privo di ambiguità (Guercio 2016). In un articolo del 2007, Maria Guercio rilevava questa complessità annoverando tra «i molteplici aspetti connessi alla conservazione digitale ... ad esempio i formati dei file, i servizi di registrazione dei formati medesimi, i modelli organizzativi per la gestione di depositi digitali», sottolineando contestualmente come la letteratura allora esistente risultasse già molto vasta ma frammentaria (Guercio 2007). Nel quindicennio successivo, gli archivisti informatici hanno dedicato molto lavoro di elaborazione teorica e pratica lungo le direttrici già delineate, focalizzandosi da una parte sulla necessità di garantire l'accessibilità e fruibilità delle risorse a lungo termine tramite l'adozione di standard e modelli in grado di prevenire l'obsolescenza di formati e applicazioni, dall'altra sull'elaborazione di procedure di trasmissione che assicurino l'autenticità e l'integrità dei documenti (Guercio 2007; Guercio 2013). Ciò ha determinato una polarizzazione intorno a temi e problemi riferibili al patrimonio archivistico, mentre sono rimasti sostanzialmente inesplorati sia in ambito accademico sia professionale terreni affini ma distinti quali la conservazione delle risorse web, delle biblioteche digitali (la cui definizione stessa rimane spesso ambigua e controversa anche nella letteratura di settore) e della documentazione bibliografica (Guercio 2016; Biagetti 2019). Manca ancora, a livello nazionale ma per lo più anche in ambito internazionale, una strategia chiara e condivisa su cosa conservare e in che modo: basti pensare ai progetti di ricerca, che sempre più frequentemente si concretizzano in prodotti fruibili sul web ma, salvo rari casi, sono eliminati dai server accademici al termine del rapporto dei responsabili con i rispettivi atenei.

Il continuo e inevitabile processo di aggiornamento delle risorse digitali, necessario per garantirne l'accesso nel corso del tempo, ha fatto affermare a Carlo Federici che «la conservazione si identifica con la salvaguardia dei materiali costituenti il bene culturale i quali supportano il testo e conferiscono ad esso quella forma che ne consente la percezione. In ogni caso

---

<sup>1</sup> <https://www.riformeistituzionali.gov.it/la-legge-costituzionale-in-materia-di-tutela-dell-ambiente/>(consultato il 5/5/2023)

ciò vige solo per gli originali – o al massimo per le riproduzioni “analogiche” – poiché *la riproduzione digitale, essendo priva di materialità, sfugge alle leggi della conservazione, tant’è che la sua trasmissione al futuro si basa sostanzialmente su procedure di copia*» (Federici 2007, corsivo mio).

Se è possibile concordare sul fatto che una risorsa digitale sia essenzialmente «un fenomeno puramente elettromagnetico, che viene reso visibile attraverso appositi software» all’interno di un ambiente unidimensionale (Orlandi 2010), non sembra tuttavia del tutto corretto considerare le risorse digitali, siano esse native o derivate da oggetti analogici, come qualcosa di ‘privo di materialità’. Quella espressa da Federici rimane comunque, esplicitamente o più spesso implicitamente, una concezione largamente condivisa delle risorse digitali. La mancanza, in molti casi, di un contatto diretto e immediato con i supporti magnetici, collocati spesso a notevole distanza fisica da chi li utilizza accedendovi tramite la rete, alimenta una visione del digitale assimilato a qualcosa di etereo. Dalla volatilità dei dati digitali (Orlandi 2010) deriva immediatamente l’idea di immaterialità. Questo assunto si combina talvolta con altri argomenti per giustificare una prospettiva in cui le risorse digitali si collocano su un piano ontologicamente diverso rispetto a quelle analogiche e tangibili. Questa dicotomia è alla base di infiniti quanto sterili dibattiti sulla presunta superiorità di un tipo di risorsa sull’altro: se da una parte trova posto la strenua resistenza dei sostenitori del libro ‘tradizionale’, dall’altra colloca il fronte positivista che attribuisce al digitale una intrinseca connotazione di ordine e pulizia, enfatizzandone il contrasto con il disordine e la sporcizia dei ‘vecchi’ supporti.

### 3. DEMATERIALIZZAZIONE O SMATERIALIZZAZIONE: UN EXCURSUS IDEOLOGICO-SEMANTICO

Su queste basi si innesta e prospera il concetto di ‘dematerializzazione’, cardine della riforma in atto per migliorare l’efficienza della Pubblica Amministrazione con l’entrata in vigore, il 1° gennaio 2006, del Codice dell’amministrazione digitale (decreto legislativo 7 marzo 2005, n. 82, ampiamente modificato dal decreto legislativo 30 dicembre 2010, n. 235). «Il tema della dematerializzazione è stato per anni oggetto di lunghi ed a volte aspri dibattiti. Oggi è balzato all’attenzione della Pubblica Amministrazione perché finalmente si vede la concreta possibilità di realizzare quegli ingenti benefici, in termini di risparmio e di efficacia, che il passaggio del documento amministrativo dalla carta al bit ha sempre promesso. Con l’entrata in vigore del Codice dell’amministrazione digitale, viene data attuazione, ricorrendo alle più avanzate tecnologie informatiche, ai meccanismi deputati a realizzare in concreto la tanto auspicata “scomparsa della carta” [...] per un percorso concreto di dematerializzazione, da un lato per eliminare i documenti cartacei attualmente esistenti negli archivi, dall’altro per ridurre significativamente la creazione di nuovi documenti cartacei. [...] Per recuperare il passato, cioè per eliminare la vecchia carta, occorrerà innanzitutto favorire l’operato delle commissioni di scarto, anche con nuove regole, improntate a pragmatismo e realismo. Nel mondo dei privati, nella maggior parte dei casi è sufficiente solo un po’ di iniziativa per ottenere grandi risparmi» (Ridolfi 2006).

Come opportunamente notato dallo stesso Ridolfi, «il termine “dematerializzazione” non possiede uno spessore semantico particolare nell’ambito amministrativo, pubblico e privato, ed è totalmente assente nella classica letteratura archivistica, mentre invece compare, anche se da poco tempo, in alcuni testi normativi e all’interno di disposizioni in materia di gestione documentale. Si tratta, tuttavia, di un termine dal forte valore evocativo per identificare la progressiva *perdita di consistenza fisica* da parte degli archivi delle amministrazioni, tradizionalmente costituiti da documenti cartacei, all’atto della loro sostituzione con registrazioni informatiche. È quindi possibile giustificare la nascita di questo termine come conseguenza diretta del progressivo incremento della gestione documentale informatizzata all’interno delle strutture amministrative» (Ridolfi 2006; corsivo mio).

Se il termine in sé si presenta dunque come un neologismo, il prefisso ‘de’ è, di contro, attestato nella nostra lingua con una gamma di significati che comprendono i concetti di allontanamento, abbassamento o movimento dall’alto verso il basso fino ad arrivare alla privazione o negazione del verbo o sostantivo a cui si antepone.<sup>2</sup> Nella letteratura di dominio è

---

<sup>2</sup> Cfr. il *Vocabolario on line* Treccani: «de- <dè> [dal lat. *de, de-*]. – 1. Prefisso, soprattutto verbale, che si trova in molte voci di derivazione latina, nelle quali indica ora allontanamento (per es. *deviare, deportare*), ora abbassamento o movimento dall’alto in basso (per es. *degradare, deprimere, declinare*), ora privazione (per es. *dedurre, detrarre*; cfr. anche *demente*), ora ha valore negativo (per es. *decrescere*), ora serve soltanto alla formazione di verbi tratti da sostantivi o aggettivi oppure, con funzione intensiva o con sign. particolari, da altri verbi (per es. *decurtare, designare, determinare*, ecc.). In verbi di formazione recente, spesso formati sull’esempio del francese, corrisponde per lo più a *dis-* o *s-* (per es. *demoralizzare, denaturare*; va notato che in francese il pref. *dé-* corrisponde non all’ital. *de-* ma a *dis-*1, essendo derivato dal lat. *dis-*, e indica separazione, privazione o azione contraria); in qualche caso serve a formare doppiotti di voci già esistenti (cfr. *demagliare, defogliare, degusciare* di fronte a *smagliare, sfogliare, sgusciare*): <https://www.treccani.it/vocabolario/de/> (consultato il 5/5/2023).



attestata anche la variante ‘smaterializzazione’<sup>3</sup>, a dimostrazione dell’assimilabilità del significato a un vero e proprio dissolvimento della materia.

Questa scelta semantica sembra volta a evocare una visione antitetica e quasi liberatoria rispetto a una quotidianità fatta di armadi e uffici polverosi, e trova compimento nella retorica che celebra con enfasi positivista «la tanto auspicata “scomparsa della (vecchia) carta”» grazie a «regole improntate a pragmatismo e realismo» nel disgraziato mondo della pubblica amministrazione, mentre «nel mondo dei privati, nella maggior parte dei casi è sufficiente solo un po’ di iniziativa per ottenere grandi risparmi» (Ridolfi 2006). L’idea di ‘pulizia’ appare infatti strettamente connessa in primo luogo con quella di ‘risparmio’ a più livelli: come miglioramento nella gestione dello spazio e delle risorse, prima di essere economico ed ergonomico; e poi, ancora, con quella di ‘ecologico’: a differenza delle vecchie scartoffie, il digitale non utilizza e non spreca carta, inchiostro né toner; non richiede francobolli né corrieri; in definitiva, dunque, si presume che non sporchi e non inquina.

Questa visione trionfalistica del digitale si rivela, a ben vedere, miope e ingannevole per l’equivoco fondamentale che il digitale sia qualcosa di completamente astratto e avulso da qualsiasi sostanza fisica. Se una digitalizzazione correttamente impostata può contribuire a liberare scrivanie e scaffali, a sveltire le procedure e a migliorare la reperibilità dei documenti, il processo rimane, di fatto, quello del trasferimento di un contenuto da un supporto a un altro di tipo diverso. L’attuazione puramente meccanica dei processi di digitalizzazione, priva dell’opportuno approfondimento metodologico-teorico e di una modellizzazione adeguata, può causare sprechi e danni non previsti risultando addirittura controproducente (Orlandi 2010).

#### 4. (ANTI)MATERIA ED ENERGIA

In linea con la visione acritica appena delineata, l’esistenza di un inquinamento derivante dall’uso del digitale è ampiamente sconosciuto e sottostimato. Tra le attività umane più impattanti sul pianeta, con emissioni di CO<sub>2</sub> e che ammontano a milioni di tonnellate annuali, si collocano le cripto valute, con una stima per il 2019 di 68 milioni di tonnellate; seguono la tecnologia in cloud e i data centre, con una stima per il 2020 di 160 milioni di tonnellate; per il complesso globale delle applicazioni e infrastrutture informatiche il totale raggiunge circa 1.4 bilioni di tonnellate (Berners-Lee 2020).

Sebbene il dato digitale in sé sia un fenomeno elettromagnetico basato sul trasporto di flussi di energia e non di materia, per il funzionamento della tecnologie digitali sono ovviamente necessari supporti fisici (magnetici) e, appunto, energia. Per quanto queste trasmettano l’informazione in modo più rapido e preciso (limitandone cioè l’entropia) rispetto ai precedenti sistemi analogici (Ciotti e Roncaglia 2006), la quantità di dati scambiati e diffusi negli ultimi decenni è cresciuta esponenzialmente, aumentando di conseguenza anche il livello di fabbisogno energetico e finendo quindi per aggravare la situazione. Questo fenomeno è conosciuto come effetto rimbalzo (*rebound effect*): in seguito all’introduzione di tecnologie più economiche ed efficienti, i benefici del risparmio ottenuto (in questo caso, di natura energetica) sono di fatto vanificati dall’incremento complessivo dei consumi che ne deriva (Berners-Lee 2020). All’apertura del convegno di Europeana 2022 – *making digital culture count*, tenutosi a L’Aia dal 28 al 30 settembre 2022<sup>4</sup>, Harry Verwayen, General Director della Europeana Foundation, ha ricordato ai presenti che, seppure semplicemente seduti in ascolto, proprio in quel momento stavano contribuendo all’innalzamento dei gas serra nell’atmosfera a causa del costante traffico dati dei propri smartphone.

#### 5. RIFLESSIONI, SPUNTI E AZIONI

In questa prospettiva, la ‘dematerializzazione’ si presenta ancora oggi, in modo più o meno consapevole, come un prodotto di *greenwashing*,<sup>5</sup> evidente, ad esempio, in un articolo eloquentemente intitolato «PEC, soluzione green: impatto ambientale ridotto e meno carta anche nella PA», in cui l’utilizzo della posta elettronica certificata (PEC) è incoraggiato per «rendere più ecologiche le procedure documentali, attraverso l’eliminazione della carta» e «ridurre l’impatto ambientale» (Pisanu 2019).

Se confrontato con la crescita della richiesta di servizi digitali, l’aumento delle emissioni di gas serra nel periodo 2015-2021 è da considerarsi nel complesso limitato soprattutto grazie al miglioramento dell’efficienza energetica, al ricorso all’energia rinnovabile e alla decarbonizzazione delle reti elettriche in atto; ciononostante, per azzerare le emissioni entro il 2050, obiettivo di Net Zero Emissions by 2050 Scenario (NZE), è necessario dimezzarle entro il 2030 (Kamiya 2022).

<sup>3</sup> Si veda per esempio Guercio 2016, p. 22.

<sup>4</sup> [https://pro.europeana.eu/files/Europeana\\_Professional/Event\\_documentation/Events/Europeana/2022/Europeana-2022-Digital-Programme.pdf](https://pro.europeana.eu/files/Europeana_Professional/Event_documentation/Events/Europeana/2022/Europeana-2022-Digital-Programme.pdf) (consultato il 5/5/2023).

<sup>5</sup> Cfr. il *Vocabolario on line* Treccani: «Strategia di comunicazione o di marketing perseguita da aziende, istituzioni, enti che presentano come ecosostenibili le proprie attività, cercando di occultarne l’impatto ambientale negativo»: [https://www.treccani.it/vocabolario/greenwashing\\_%28Neologismi%29/](https://www.treccani.it/vocabolario/greenwashing_%28Neologismi%29/) (consultato il 5/5/2023).



	2015	2021	Change
Internet users	3 billion	4.9 billion	60%
Internet traffic	0.6 ZB	3.4 ZB	440%
Data centre workloads	180 million	650 million	260%
Data centre energy use (excluding crypto)	200 TWh	220-320 TWh	9.4
Crypto mining energy use	4 TWh	100-140 TWh	+2 300-3 300%
Data transmission network energy use	220 TWh	260-340 TWh	19.4

Tabella 1. Global trends in digital and energy indicators, 2015-2021 (fonte: Kamiya 2022).

Una nuova consapevolezza e l'agire conseguente devono passare attraverso il riesame di convinzioni maturate in modo acritico. Un chiaro e documentato prospetto di quali azioni abbiano un impatto ambientale, corredato da strumenti che mettano il lettore in grado di eseguire in autonomia calcoli approssimativi analoghi riferiti a qualsiasi attività e impegnarsi concretamente nel cambiamento, è fornito dall'opera di Mike Berners-Lee *How bad are bananas? The carbon footprint of everything*, alla cui prima edizione del 2010 è seguita, dieci anni dopo, una seconda profondamente rivista e aggiornata (Berners-Lee 2020).

All'interno di Europeana, una fondazione internazionale sorta per iniziativa dell'Unione europea e finanziata dalla stessa Unione europea e dagli Stati membri,<sup>6</sup> si è formata recentemente la Climate Action Community, una comunità aperta a tutti che propone specificamente di approfondire la ricerca, portare avanti la divulgazione e impegnarsi pacificamente nelle battaglie su queste tematiche.<sup>7</sup> Alla pubblicazione dell'Europeana Climate Action Manifesto nel 2021<sup>8</sup> sono seguite varie iniziative da parte dello Steering Group, come l'organizzazione di momenti di dialogo e confronto di informazione e sensibilizzazione. Tra le attività in corso si evidenzia un lavoro di indagine condotto da un gruppo di lavoro (Environmental Sustainability Practice Task Force) appositamente costituito all'interno dello Steering Group con la collaborazione di esperti esterni e guidato da Evangelia Paschalidou, impegnata nella ricerca sulla sostenibilità ambientale delle infrastrutture informatiche per la conservazione del patrimonio culturale digitale. L'obiettivo è quello di delineare le pratiche di gestione e conservazione dei dati all'interno delle istituzioni culturali come punto di partenza per elaborare una strategia per una trasformazione digitale ecologicamente sostenibile.<sup>9</sup> Sullo sfondo di un panorama disperso e frammentario, l'iniziativa in questo campo da parte di un ente finanziato dall'Unione Europea appare particolarmente significativa.

In ambito nazionale, il contesto normativo pertinente per lo sviluppo ecosostenibile nell'ambito dei beni culturali è rappresentato dal Piano Nazionale di Digitalizzazione del patrimonio culturale (PND) predisposto dall'Istituto centrale per la digitalizzazione del patrimonio culturale – *Digital Library* del Ministero della cultura nel 2022, frutto di un processo di condivisione e confronto con diverse istituzioni culturali e che si propone come riferimento metodologico e operativo per tutte le istituzioni e gli operatori culturali, sia in ambito pubblico che privato, che si riconoscono nei valori in esso enunciati: «Il PND costituisce la visione strategica con la quale il Ministero ... intende promuovere e organizzare il processo di trasformazione digitale nel quinquennio 2022-2026 nei diversi settori dell'ecosistema culturale, rivolgendosi in prima istanza ai musei, agli archivi, alle biblioteche, alle soprintendenze, agli istituti e ai luoghi della cultura pubblici che conservano, tutelano, gestiscono e/o valorizzano beni culturali» e «costituisce anche il contesto strategico, intellettuale e tecnico di riferimento per la realizzazione degli obiettivi del Piano Nazionale di Ripresa e Resilienza (PNRR) per l'investimento MIC3 1.1 Strategie e piattaforme digitali per il patrimonio culturale».<sup>10</sup>

## BIBLIOGRAFIA

[1] Berners-Lee, Mike. *How bad are bananas? The carbon footprint of everything*. Revised 2000 edition. London: Profile Books Ltd, 2020.

[2] Biagetti, Maria Teresa. *Le biblioteche digitali: tipologie, funzionalità e modelli di sviluppo*. Milano: FrancoAngeli, 2019.

<sup>6</sup> <https://www.europeana.eu/it/about-us> (consultato il 5/5/2023).

<sup>7</sup> <https://pro.europeana.eu/page/climate-action-community> (consultato il 5/5/2023).

<sup>8</sup> <https://pro.europeana.eu/post/the-europeana-climate-action-manifesto> (consultato il 5/5/2023).

<sup>9</sup> Europeana Climate Action Community Work Plan 2022-2023, disponibile online: <https://pro.europeana.eu/post/climate-action-community-work-plan-2022-2023> (consultato il 5/5/2023), p. 6: «Digital Information Management Survey – design, implementation and analysis. Led by the Environmental sustainability practice Task Force, we will launch a survey on data management and preservation practices in recognition that there is currently no clear overview of how cultural heritage institutions and individuals plan for both a green and digital transformation. We will solicit expert advice if and when needed. This survey is a formative action that supports Europeana's Climate Action Manifesto. It will advocate embedding optimal working practices that minimise the digital cultural heritage sector's impact on the climate and environment. The survey could be followed by identifying and establishing a common vocabulary; clustering of mapping tools and performance indicators, a compilation and comparison of climate action institutional manifestos to develop an informative overview of the sector.»

<sup>10</sup> <https://digitallibrary.cultura.gov.it/il-piano/> (consultato il 5/5/2023).

- [3] Ciotti, Fabio e Gino Roncaglia. Il mondo digitale: introduzione ai nuovi media. Roma-Bari: Laterza, 2006.
- [4] The Climate Heritage Network 2022-24 Action Plan – Empowering People to Imagine and Realise Climate Resilient Futures Through Culture – from Arts to Heritage: <https://www.climateheritage.org/actionplan> (consultato il 5/5/2023).
- [5] La dematerializzazione della documentazione amministrativa. Numero monografico «i Quaderni» 24 (aprile 2006), supplemento a «InnovAzione», 9 (2006).
- [6] The Europeana climate action manifesto: <https://pro.europeana.eu/post/the-europeana-climate-action-manifesto>(consultato il 5/5/2023).
- [7] Federici, Carlo.«La conservazione dei materiali librari.» Giovanni Solimine e Paul Gabriele Weston (a cura di). Biblioteconomia: principi e questioni: 379-394. Roma: Carocci, 2007.
- [8] Guercio, Maria.«La conservazione delle memorie digitali.» Giovanni Solimine e Paul Gabriele Weston (a cura di). Biblioteconomia: principi e questioni: 395-412. Roma: Carocci, 2007.
- [9] Guercio, Maria.Conservare il digitale. Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali.Roma-Bari: Laterza, 2013.
- [10] Guercio, Maria.« Conservare il digitale: modello nazionale e contesto internazionale.» DigitCult1 (2016), 2: 19-26. <http://dx.doi.org/10.4399/97888548960932> (consultato il 5/5/2023).
- [11] InterPARES Project Terminology Database. [http://www.interpares.org/ip2/ip2\\_terminology\\_db.cfm](http://www.interpares.org/ip2/ip2_terminology_db.cfm) (consultato il 5/5/2023).
- [12] Kamiya, George.Data Centres and Data Transmission Networks.Tracking report – September2022.International Energy Agency (IEA). <https://www.iea.org/reports/data-centres-and-data-transmission-networks> (consultato il 5/5/2023).
- [13] Kilbride,William. «Memories from the Anthropocene: digital preservation in a time of climate crisis». Digital Preservation Coalition. <https://www.dpconline.org/blog/memories-from-the-anthropocene-digital-preservation-in-a-time-of-climate-crisis> (consultato il 5/5/2023).
- [14] Marzano, Gilberto.Conservare il digitale.Metodi, norme, tecnologie.Milano: Bibliografica, 2011.
- [15] Ministero della Cultura (MiC). Istituto centrale per la digitalizzazione del patrimonio culturale – Digital Library. Piano Nazionale di Digitalizzazione del patrimonio culturale 2022-2023. Versione 1.1. [https://digitallibrary.cultura.gov.it/wp-content/uploads/2023/01/PND\\_versione1\\_1\\_gen2023.pdf](https://digitallibrary.cultura.gov.it/wp-content/uploads/2023/01/PND_versione1_1_gen2023.pdf) (consultato il 5/5/2023).
- [16] Orlandi, Tito. Informatica testuale: teoria e prassi. Roma-Bari: Laterza, 2010.
- [17] Pisanu, Nicoletta.«PEC, soluzione green: impatto ambientale ridotto e meno carta anche nella PA». Agenda Digitale,13/11/2019, online: <https://www.agendadigitale.eu/documenti/pec-soluzione-green-impatto-ambientale-ridotto-e-meno-carta-anche-nella-pa/> (consultato il 5/5/2023).
- [18] Pontevolpe,Gianfranco. «Gli obiettivi del Governo italiano per la dematerializzazione dei documenti.» Stefano Pigliapoco (a cura di). Conservare il digitale: 15-30. Macerata: eum, 2010.
- [19] Ridolfi, Pierluigi.«Dematerializzazione dei documenti: idee per un percorso.» Ministro per l’Innovazione e le Tecnologie. La dematerializzazione della documentazione amministrativa. Libro Bianco del Gruppo di Lavoro interministeriale per la dematerializzazione della documentazione tramite supporto digitale: 5-20. Roma: CNIPA, 2006.

# New Media Art: indagini conservative sulle eredità immateriali dell'arte di fine millennio

Lorenza Carannante<sup>1</sup>

<sup>1</sup> Università degli Studi della Campania Luigi Vanvitelli, Italia – lorenza.carannante@unicampania.it

## ABSTRACT

L'obsolescenza incalzante delle eredità digitali non lascia margini d'intervento, imponendo un'attività conservativa e di restauro puntuale e sistematica. Lo studio sul restauro dell'arte New Media ha raggiunto vertici internazionali, dando vita ad un interesse attivo volto all'approfondimento e alla messa a punto di metodologie precise. Ad oggi, Media e New Media Art potrebbero essere considerate una sorta di risposta tecnologica all'Arte Concettuale, in cui il *medium*, che molte volte è opera stessa, proviene da un linguaggio nuovo e in continuo movimento che non ha nulla a che vedere con l'arte tradizionale, e l'idea, pur essendo spesso del tutto celata, è messaggio fondante dell'intera operazione artistica. La natura del supporto della contemporaneità digitale è estremamente multiforme e pertanto soggetta a leggi temporali e conseguenti approcci conservativi assolutamente diversi tra loro, tali da richiedere un aggiornamento costante degli strumenti e procedimenti che spesso impongono la riscrittura, se non una vera e propria elaborazione *ex novo*, di regole e criteri metodologici; il relativo approfondimento sarà indagato procedendo prima attraverso un'analisi di tipo strettamente teorico sullo stato dell'arte, procedendo poi con un'indagine condotta su ipotetici metodi conservativi sui quali sarà utile soffermarsi in conclusione per considerare l'ampiezza e l'eterogeneità di informazioni che possano essere di riferimento fondamentale in ambito teorico. Da un punto di vista metodologico, risulta utile il perseguimento di una rilettura, eventuale rielaborazione e successiva possibilità di applicazione di alcuni aspetti delle teorie del passato che possano fungere da guida iniziale, nonché il successivo approfondimento delle proposte recenti sulla conservazione e il restauro contemporaneo, affinché si possa arrivare ad una regola nuova, inedita e soprattutto condivisibile in merito alla New Media Art, partendo anzitutto dal necessario confronto con la tradizione.

## PAROLE CHIAVE

Media Art; New Media Art; Conservation; Preservation; Cultural heritage

## 1. NEW MEDIA ART: INDAGINI CONSERVATIVE SULLE EREDITÀ IMMATERIALI DELL'ARTE DI FINE MILLENNIO

Lo studio relativo alla conservazione e al restauro dell'arte New Media, oltre alle relative indagini e attività di ricerca svolte caso per caso, ha raggiunto vertici a livello internazionale, dando vita ad un interesse attivo volto all'approfondimento e alla messa a punto di metodologie precise e sistematiche. Numerose le pubblicazioni italiane in merito, basti pensare agli studi più recenti condotti da Valentino Catricalà e Domenico Quaranta, ad esempio, nonché gli incontri accademici che hanno instradato l'attenzione su necessità evidenti, oltre che inevitabili, guardando ad un futuro del restauro di materiali "immateriali" rispetto ai quali non si è ancora totalmente preparati poiché in continua, incessante evoluzione. Nuovi gli attori coinvolti all'interno della disciplina, ormai aperta a personalità ancora altre rispetto a quelle chiamate a rispondere già a partire dagli anni '60 grazie alle peculiarità dell'Arte Concettuale, come ingegneri e tecnici del suono, ad esempio, e ancora informatici, programmatori, progettisti e sviluppatori di componenti hardware e software. Ad oggi, Media e New Media Art, le arti elettroniche più in generale, la net.art o ancora la Digital Art, per citarne alcune, nell'interezza della loro complessità già a partire da una definizione plausibile che aiuti gli addetti ai lavori – e non solo – a circoscriverle in un determinato approccio tecnico-stilistico, potrebbero essere considerate come una sorta di risposta tecnologica alla sopracitata Arte Concettuale, in cui il *medium*, che molte volte è formalmente opera stessa, ricoprendo al contempo necessità iconografiche ed iconologiche, proviene da un linguaggio nuovo, inedito, in continuo movimento e che, soprattutto, non ha nulla a che vedere con l'arte tradizionale, e l'idea, pur essendo spesso del tutto celata, è messaggio fondante dell'operazione artistica tutta. La natura del supporto tecnologico è estremamente multiforme e pertanto soggetta a leggi temporali inevitabili e conseguenti, oltre che urgenti, approcci conservativi assolutamente diversi tra loro in via di necessaria e definitiva messa a punto. Il nuovo scenario cui gli operatori sono davanti è fluido e mai uguale. Scrive Lev Manovich:

I vecchi media implicavano un creatore, che assemblava manualmente gli elementi testuali, visivi e/o sonori in una determinata composizione o sequenza. Quella sequenza veniva poi immagazzinata in un certo formato, in un ordine fisso e imm modificabile. [...] I nuovi media, invece, sono caratterizzati dalla variabilità. Altri termini, usati spesso in relazione ai nuovi media, e che potrebbero essere sinonimi di *variabile*, sono *mutabile* e *liquido*. Invece di riprodurre tante copie identiche, un nuovo oggetto mediale riproduce tante versioni diverse e, invece di essere create integralmente da un essere umano, queste versioni vengono spesso assemblate da un computer.<sup>1</sup>

In relazione a quest'ultimo *medium* citato, ne viene proposta una definizione dagli informatici Alan Kay e Adele Goldberg, che ne introdussero anche una teoria sul finire degli anni '70 con il concetto di *metamedium*<sup>2</sup>, considerando il macchinario, quindi, come uno strumento capace di inglobare e conseguentemente simulare altri strumenti, permettendone così la creazione di nuovi, come anche nuovi linguaggi artistici, ad esempio, forme ed estetiche inedite, ma che in un certo qual modo rispondono a principi e problematiche non del tutto dissimili da quelli dell'arte contemporanea, nonostante gli ambiti di ricerca sia teorica che pratica dedicati alle indagini di tutela, conservazione e restauro siano stati sottoposti, nel tempo, a una nuova tipologia di azione ancora in fase embrionale legata alla preservazione di questi strumenti.

Negli anni, sono stati e sono numerosi, oltre che puntuali, come accennato, gli interventi di studiosi italiani, com'è il caso, ad esempio, di Laura Barreca, a sostegno anzitutto della documentazione fornita al conservatore da parte dell'artista – che è anche programmatore, rispetto alla sua opera e ai materiali utilizzati<sup>3</sup>; o ancora quello di Marco Mancuso, che ha monitorato l'evoluzione e l'impatto delle tecnologie e della scienza sull'arte attraverso l'osservatorio online Digicult, creando un archivio ad oggi imprescindibile per completezza di fonti. In generale, comunque, si è sempre privilegiata la conservazione delle componenti immateriali e concettuali rispetto a quelle fisiche, condannate, com'è immaginabile da quanto detto finora, ad una illeggibilità dipendente da un'obsolescenza repentina inevitabile.

A livello internazionale, invece, il dibattito si è fortemente consolidato già a partire dalla fine degli anni '80 con l'interesse dimostrato dai numerosi enti competenti – strutture fisiche e piattaforme digitali in rete, i quali s'impegnano tuttora nell'indagine conservativa e di archiviazione digitale, coinvolgendo gli esperti di ciascun settore utile, conferendo inoltre medesimo valore artistico sia agli strumenti utilizzati che al concetto celato all'interno dell'opera di riferimento, scelta emblematica da non sottovalutare poiché precedentemente spesso ignorata. È il caso, ad esempio, del lavoro costante svolto dal centro di ricerca ZKM di Karlsruhe, dei progetti curati dalla piattaforma Rizome.org, o, ancora, di conferenze che ricoprono un ruolo d'importanza imprescindibile come *Modern Art: Who Cares?* del 1997, organizzata e discussa dalla Foundation for the Conservation of Modern Art di Amsterdam, i cui atti, poi pubblicati, sono ad oggi di grande riferimento. L'artista tende ad utilizzare e fare propri tutti gli elementi e gli strumenti che la sua contemporaneità mette a disposizione. La ricerca artistica, com'è desumibile, è e sarà sempre permeata quindi da un forte senso di sperimentazione rispetto alle evoluzioni industriali legate al suo tempo, cosa che andrà di pari passo con un disinteresse sempre più forte, sebbene inconscio, rispetto alle sorti dei materiali adottati. Oltre a questo assunto che, per quanto breve, risulta essere di grande importanza per inquadrare il cambiamento che in qualche modo contribuisce a spezzare l'evoluzione più o meno lineare avviata dalla tradizione, è utile specificare che durante tutto il Novecento l'arte vive una serie di sviluppi e di avanzamenti anche, e soprattutto, da un punto di vista tecnologico: il passaggio dal *medium* analogico a quello digitale non è, infatti, da considerarsi di poco conto, in particolare se si guarda al futuro della produzione artistica che ne deriva, al di là della nascita di nuove strategie legate ad ambiti come la curatela, gli allestimenti museali, il mercato dell'arte, per riportare alcuni esempi, e alla stessa volontà di storicizzare il frutto dell'inarrestabile progresso tecnologico. Sociologicamente, come sostenuto da McLuhan<sup>4</sup>, il *medium* è un'estensione dell'essere umano, sia esso analogico che, come si è visto, digitale: come accennato, ciascun tempo, qualsiasi fase della comunicazione ha goduto della rivoluzione espressiva offerta dal *medium* di riferimento. Ogni strumento ha una sua contemporaneità e a suo modo è stato necessario ad avviare ed avviare una consapevolezza estetica altrimenti potenzialmente inesprimibile. Mutando nel corso del tempo, l'obiettivo dell'artista non è più legato all'imitazione perfetta e fedele della natura circostante, ma all'intuizione dipendente dall'adozione ed il successivo utilizzo di strumentazioni indipendenti da ciò che la tradizione aveva loro concesso fino ad allora, in una visione nuova del rapporto uomo, tempo ed arte, in cui il prodotto artistico non ricopre più il ruolo di veicolo di rappresentazione storica della realtà, ma vuole esserne esperienza attraverso gli sviluppi tecnologici, il costante mutamento della materia "immateriale", liquida, dei supporti digitali. Scrive Valentino Catricalà, in qualche modo accostando la rottura tecnico-stilistica introdotta dalle avanguardie storiche a quella dell'arte Media e New Media:

<sup>1</sup>Manovich, Lev. Il linguaggio dei nuovi media. Edizioni Olivares, 2002, p. 57.

<sup>2</sup>Ibidem.

<sup>3</sup>Mancuso, Marco. Arte, tecnologia e scienza. Le art industries e i nuovi paradigmi di produzione nella new media art contemporanea. Mimesis Edizioni, 2018, p. 48.

<sup>4</sup>McLuhan, Marshall. Gli strumenti del comunicare. Il saggiautore, 2005, p. 90.

L'importanza delle avanguardie storiche, come è noto, risiede nella coppia dialettica rottura/innovazione: una rottura con il passato a fronte di programmi e prospettive per il futuro. Una coppia, questa, che non si distacca da una struttura temporale fedele a una idea di progresso. [...] Una rottura con il passato e, dunque, con i linguaggi tecnici dell'arte classica, delle cosiddette "belle arti". Negazione dell'esistente, dunque, a favore di una ricostruzione dei linguaggi e delle modalità di creazione dell'esistente stesso. [...] È noto, inoltre, che all'interno di questa dialettica emerge una ristrutturazione dei linguaggi dell'arte operata anche grazie all'inglobamento di materiali fuori dall'ambito ritenuto prettamente artistico.<sup>5</sup>

Inoltre, come posto all'attenzione da Domenico Quaranta<sup>6</sup>, la produzione artistica Media e New Media, così come le espressioni comunicative successive, fanno riferimento ad uno spettro ben preciso e molto più ampio degli strumenti a disposizione per una libertà espressiva totale e totalizzante. Termini, a ben vedere, già precedentemente presenti e attivamente utilizzati nel tentativo di introdurre criticamente una tipologia inedita di fare artistico, come si diceva, "immateriale", guardando alla materia tradizionalmente intesa. È come se questo particolare movimento, quello legato alle tecnologie elettroniche anzitutto, per poi abbracciare definitivamente i suoi successivi avanzamenti digitali, si sviluppasse in una nuova cornice delle arti inedita a tal punto da rendere ostico anche solo il tentativo di definirne le peculiarità partendo dai termini che ne compongono il nome: Media Art, New Media Art, in cui riconosciamo opere computer-based, digitali, online, network-based, per citarne alcune tipologie. La numerosità dei diversi sviluppi di questa tipologia di fare artistico ha contribuito, quindi, a rendere quasi del tutto impossibile una definizione teorica che fosse univoca e stabile nel tempo, guardando soprattutto ai classici e alla loro necessità costante di trovare nell'arte del passato un linguaggio di riferimento tale da fungere da guida per gli approcci successivi. Riportando le parole di Hanna Barbara Hölling:

The term computer-based art is wide and difficult to narrow to a single definition. In a time of enormous expansion of new media and the perpetual flux of encoded information on the Internet, almost every art installation involves computation at a certain stage of its development and/or presentation. [...] As a rule, computer-based artworks consist of files, an operating system, software, and hardware. [...] In contrast to a large portion of traditional fine art, where the point of departure for the assessment of its status lies in defining the authenticity or originality of the artwork in question, the method for describing computer based artworks is built on the complex term-integrity. The integrity of computer-based installations depends on hardware and software components. Conceptual integrity refers to the work's relationship to the process or technology employed, and the social and cultural setting in which the work was created.<sup>7</sup>

La rivoluzione digitale, in questo caso quindi circoscritta alla produzione artistica, si fa sempre più spazio e procede per forza di cose, anche se implicitamente, con l'evoluzione di strategie e sfide in campo conservativo, con aree focali destinate a cambiare e, di conseguenza, ad essere messe in discussione in relazione alla loro durabilità ed eventuale permanenza. Sebbene affidarsi alle pratiche maggiormente accolte allo stato attuale e agli standard stabiliti possa effettivamente avere i suoi riscontri positivi, non è da considerarsi una strategia infallibile poiché, come si è accennato, le sfumature sono innumerevoli e il discorso è tanto potente quanto labile a causa dell'infinità di casistiche differenti che lo compongono. Da un punto di vista conservativo – tenendo in considerazione anche l'ambito preventivo – e restaurativo, nonché, come stabilito all'interno del "Variable Media approach" proposto dal Variable Media Network<sup>8</sup>, del Deposito (la conservazione tradizionale di un'opera in via fisica attraverso dispositivi dedicati alla memorizzazione o all'archiviazione digitale di file su disco), della Emulazione (l'imitazione dell'aspetto originario dell'opera su un supporto completamente diverso), della Migrazione (l'utilizzo di tecnologie superiori che permettono di trasportare l'opera da un supporto a un altro), e della Reinterpretazione (propriamente, la reinterpretazione dell'opera ogni volta che deve essere reinstallata), si è di fronte ad un argomento indubbiamente complesso che richiede la costruzione di una strategia specifica. Tale strategia non riguarda

<sup>5</sup>Catricalà, Valentino. *Media Art. Prospettive delle arti verso il XXI secolo. Storie, teorie, preservazione*. Mimesis Edizioni, 2016, pp. 39-40

<sup>6</sup>Quaranta, Domenico. *Media, New Media, Postmedia*. Postmedia books, 2010, p. 27, di cui si cita testualmente: Proprio alle avanguardie storiche si ricollegano quelle esperienze che vogliono andare oltre quello che appare, allora, il vicolo cieco della stagione informale: New Dada, Nouveau Réalisme, Gutai, Happening, Fluxus, arte cinetica e programmata, Optical Art e Pop Art, Video Art. La realtà, come oggetto reale o rappresentato, entra a far parte dell'opera; la cultura pop veicolata dai media diventa oggetto dell'attenzione degli artisti; l'arte si appropria di tutti i media, dal corpo all'oggetto di consumo, dalla pubblicità al televisore alla macchina; sviluppi teorici come la cibernetica e la teoria dell'informazione informano il lessico degli artisti.

<sup>7</sup>Hölling, Hanna Barbara. *Versions, variations, and vulnerability: ethical considerations and conservation options for computer-based art*, in *The Electric Media Review*, Vol. Two (2013), in <https://resources.culturalheritage.org/emg-review/volume-two-2011-2012/versions-variations-and-variability-ethical-considerations-and-conservation-options-for-computer-based-art/>, p. 35.

<sup>8</sup>Variable Media Network website: <https://variablemedia.net/e/index.html>

esclusivamente il dato estetico, ma deve prevedere altre tipologie di ricerca, di diagnostica e di cura affinché si possa accedere ad una serie di informazioni – già digitalmente depositate e non – tale da poter conseguentemente pensare di procedere con la messa a punto di un modus operandi puntuale, sistematico e affine alle necessità di ciascuna opera d'arte. Inizialmente la Migrazione era la strategia di conservazione utilizzata con maggiore frequenza, per poi lasciare spazio anche ai principi dell'Emulazione che recentemente hanno suscitato molto più interesse rispetto al passato, sebbene non siano riconosciuti come una soluzione definitiva in relazione alla manchevolezza sia estetica che, conseguentemente, funzionale del supporto rispetto a quello originale; nozione non da poco conto poiché in qualsiasi caso la conservazione dovrebbe garantire quantomeno la possibilità di replica e riproduzione non soltanto del materiale depositario sensibile al rapido deperimento, ma anche dell'esperienza simbolica celata al suo interno, oseremmo dire invisibile, ma non per questo scevra dall'essere tramandata a letture future. Scrive in nota Giovanni Michetti:

Beninteso, la storia delle successive migrazioni da un supporto all'altro deve essere conservata, perché ogni supporto – e ogni sistema logico entro il quale l'oggetto è immerso – ha delle specifiche caratteristiche. Il rigore filologico impone di tenere traccia di questo “percorso”, non foss'altro perché il supporto stesso ci comunica informazione [...] non è affatto improbabile che un file, memorizzato in origine su un floppy disk, giunga fino a noi sotto forma di file archiviato su disco fisso. Tale circostanza non sembra problematica: la conoscenza del supporto originale serve certo per immaginare le modalità di fruizione della risorsa; e la conoscenza di un'avvenuta trasformazione deve renderci consapevoli della possibilità di perdita di dati rilevanti. [...] Ci ritroveremmo (sempre che sia stato conservato tutto l'apparato tecnologico necessario e che si possa accedere all'oggetto senza difficoltà) davanti a uno schermo e alla rappresentazione dell'oggetto.<sup>9</sup>

A rigor di logica va aggiunto e sottolineato, infine, che ogni sfida conservativa – che non si limiti a preservarne soltanto l'eventuale archiviazione digitalizzata, ad esempio, ma anche a fornirne l'accesso materiale tramite aggiornamento hardware e software, nonché l'adozione di un rigore filologico che tenga traccia di ogni avvenuta migrazione – deve affrontare non solo la tutela e la conservazione delle tecnologie del passato, come si accennava, ma anche quelle del futuro, in una fretta angosciante causata dalla durata sempre più breve dei materiali digitali, rendendo definitivamente urgente la necessità di un'azione conservativa tempestiva oltre che, naturalmente, efficace. Numerose sono le domande che sorgono spontanee a riguardo. Anzitutto, quale potrebbe essere la modalità iconograficamente consona e fedele per riattualizzare un'installazione realizzata con un hardware di una generazione ormai superata e quindi obsoleta, letteralmente; quali potrebbero essere considerate le parti – sostituibili e non – che possono ricoprire un ruolo estetico, oltre che funzionale, in una dimensione di necessità legate all'originalità, all'autenticità formale dell'opera. Domande, queste, che spingono e spostano l'attenzione verso una problematica intrinsecamente legata alla conservazione della produzione contemporanea, ovvero i problemi etici, le cui peculiarità accompagnano e spesso condizionano le scelte pratiche e viceversa. A suo modo, è plausibile pensare che la tradizione del restauro possa intervenire, accettando di essere interpretata come strategia interdisciplinare su misura e applicata in linea con le necessità sopracitate. Partendo dall'assunto secondo cui l'opera d'arte instaura di volta in volta relazioni diverse in base alle mutevoli condizioni ambientali e temporali<sup>10</sup> in cui è inserita e dalle quali in un certo qual modo dipende, bisogna considerare l'azione conservativa e di restauro come un'operazione attivamente critica, soprattutto in relazione alla teoria brandiana, rivolta quindi alla tradizione, secondo cui “il restauro costituisce il momento metodologico del riconoscimento dell'opera d'arte, nella sua consistenza fisica e nella sua duplice polarità estetica e storica, in vista della sua trasmissione al futuro”<sup>11</sup>.

Primariamente, questa asserzione sottolinea l'importanza del conferimento della dignità artistica, oltre che intellettuale, all'opera d'arte, poiché il *medium* plasmato ad oggetto artistico diviene tale solo quando viene chiamato a compiere attivamente una sperimentazione estetica, unico lasciapassare per una successiva indagine legata al suo restauro. È altresì evidente che l'assioma riguardante la consistenza fisica dell'opera d'arte, poi, richiami il primo assunto della Teoria secondo cui “si restaura solo la materia dell'opera d'arte”<sup>12</sup>, per il quale si interviene sull'aspetto materico in quanto luogo indiscusso della manifestazione dell'immagine, dell'espressione iconografica attuata formalmente<sup>13</sup>. L'intervento restaurativo in ambito contemporaneo risponde ad una serie di possibilità – ed impossibilità – derivanti, com'è desumibile,

<sup>9</sup> Michetti, Giovanni. Il paradosso della conservazione digitale: riflessioni sull'autenticità, 2010, p. 43, in <http://digitalia.sbn.it/article/view/237>

<sup>10</sup> Hölling, Hanna Barbara. Versions, variations, and vulnerability: ethical considerations and conservation options for computer-based art, in *The Electric Media Review*, Vol. Two (2013), in <https://resources.culturalheritage.org/emg-review/volume-two-2011-2012/versions-variations-and-variability-ethical-considerations-and-conservation-options-for-computer-based-art/>, p. 43.

<sup>11</sup> Brandi, Cesare. *Teoria del restauro*. Piccola Biblioteca Einaudi, 2000, p. 6.

<sup>12</sup> *Ibidem*.

<sup>13</sup> Purgar, Kresimir. *Iconologia e cultura visuale*. W. J. T. Mitchell, storia e metodo dei visual studies. Carocci Editore, 2020, p. 107.



anche dall'influenza del passato, approfondendone gli aspetti teorici prima di concretizzare quelli pratici con interventi altrimenti potenzialmente fatali per la fruizione futura dell'opera. Un'affinità e vicinanza, quella con la tradizione, che la pratica artistica della contemporaneità ha assimilato soprattutto procedendo per opposti, ponderando differenze basate sul superamento di qualsiasi limite materiale, tecnico o intellettuale, sull'innovazione digitale, come accennato, sulla volontà di decontestualizzazione nonché smaterializzazione come ulteriore, se non definitiva, forma di linguaggio. Qualora la materia costituente fosse esclusivamente mezzo espressivo, il restauratore s'impegnerà ad avviare un intervento che non ne vanifichi il significato grazie ad un precedente approfondimento sulla comprensione della natura dell'oggetto *medium*. Lo stesso Brandi, infine, sottolineando l'aspetto indispensabile della materia dell'opera in quanto rivelatrice assoluta<sup>14</sup>, impone l'esigenza di volgere l'attenzione alla possibilità o meno di considerare attuali determinati punti salienti della sua *Teoria*. Ciascun restauratore vi riconoscerà sempre un ruolo imprescindibile, sia condividendone ogni punto, sia ponendosi in una posizione critica. Se prima l'arte raccontava i grandi eventi, i cambiamenti della storia attraverso mezzi che ad oggi, a rigor di logica, vengono definiti tradizionali, ora pone in prima linea lo scopo espressivo, esasperando la sperimentazione dei procedimenti, dei *medium* time-based a disposizione che talvolta coincidono con l'opera stessa. Sarà utile riprendere un aspetto della *Teoria* brandiana per coniugarla a vantaggio della contemporaneità: Brandi sostiene, infatti, che bisogna considerare la sua *Teoria* come prerogativa applicabile ad una tipologia di opera d'arte originata non in un ambito strettamente pratico, quanto anche, e soprattutto, filosofico ed estetico, aspetto verosimilmente associabile alla produzione artistica contemporanea ed espresso anche nelle opere time-based. Ciò che non è da sottovalutare, però, è che, nonostante quest'ultima definizione possa sembrare maggiormente inclusiva rispetto alle opere d'arte tutte, la realtà della conservazione tradizionale è un'altra: limitata alla conoscenza dell'oggetto materiale e della sua natura attraverso analisi scientifiche, l'attività di tutela non s'interroga tanto sulla condizione estetica, spesso effimera, dell'opera, quanto su quella strettamente tangibile. Rimandando la conservazione dell'esperienza concettuale, che è protagonista della contemporaneità, queste dinamiche non riescono ancora ad essere esaurientemente utili, avendo quindi sempre bisogno di essere accompagnate da una forma analitica altra, rivolta principalmente al dato etico, che ne garantisca il funzionamento iconologico nel poter assicurare altresì alle generazioni future la possibilità di attingere alle eredità immateriali del nostro tempo, nella fattispecie alle operazioni digitali basate sul tempo, posizionate fenomenologicamente tra l'estetica effimera delle installazioni di metà secolo e le arti performative dipendenti dal *qui e ora*, in cui il significante non è più veicolato dalla concezione classica di arte come prodotto del genio artistico coniugato alle capacità manuali, ma anche, e soprattutto, dalla presenza imprescindibile di *medium* spesso intangibili e comunicativamente rivoluzionari.

In conclusione, partendo da un'indagine teorica e un tentativo di analisi che continui a porsi come obiettivo primario quello di rapportarsi ad una tipologia di produzione artistica carica dell'idea di obsolescenza già all'interno del suo supporto, che molte volte è opera stessa, tra i risultati cui ha auspicato lo studio proposto c'è indubbiamente la volontà di stabilire, per quanto possibile e in linea con l'inevitabilità dei limiti legati ad ogni aspetto teorico vagliato, una metodologia d'indagine e successivo intervento di conservazione preventiva e di restauro che rispetti anzitutto la nozione visivamente tangibile ed effettiva di condizione originaria<sup>15</sup> dell'opera d'arte time-based. Il periodo predeterminato di obsolescenza, la data di scadenza di ciascun *medium* prodotto industrialmente non può essere ignorato in alcun modo. La necessaria operazione di imposizione della verità, della realtà ormai passata ma indispensabile al presente del prodotto artistico, fedele alla sua forma annunciata ed espressa attraverso qualsiasi *medium* l'artista decida di utilizzare, andrebbe di pari passo con la narrazione irrinunciabile dettata in modo non sempre palese. Questo sotteso bisogno risulta quanto mai inscindibile dalla sua iconografia, quindi, non più legata all'idea tradizionale del termine, ma resistente alla concretezza di un invecchiamento intransigente che può e, anzi, deve essere affrontata con sistematicità attraverso aggiornamenti puntuali di software, recuperando gli hardware almeno in via formale in base al caso di riferimento, nonché sostituzioni, oltre alla preventiva archiviazione digitale in rapporto non soltanto alla memorizzazione del dato formalmente inteso, ma anche, e soprattutto, all'idea celata, al concetto ancor più immateriale di alcuni media, con l'utilizzo di strumentazioni, rispondenti al medesimo degrado repentino, all'interno di laboratori specializzati che ricoprono un ruolo d'importanza insostituibile.

## BIBLIOGRAFIA

- [1] Brandi, Cesare. *Teoria del restauro*. Piccola Biblioteca Einaudi, 2000.
- [2] Catricalà, Valentino. *Media Art. Prospettive delle arti verso il XXI secolo*. Storie, teorie, preservazione. Mimesis Edizioni, 2016.
- [3] Hölling, Hanna Barbara. *Versions, variations, and vulnerability: ethical considerations and conservation options for computer-based art*, in *The Electric Media Review*, Vol. Two (2013), in <https://resources.culturalheritage.org/emg-review/volume-two-2011-2012/versions-variations-and-variability-ethical-considerations-and-conservation-options-for-computer-based-art/>

---

<sup>14</sup> Brandi, Cesare. *Teoria del restauro*. Piccola Biblioteca Einaudi, 2000, p. 4.

<sup>15</sup> Muñoz Viñas, Salvador. *Teoria contemporanea del restauro*. Castelvocchi, 2017, p. 65.



- [4] Mancuso, Marco. Arte, tecnologia e scienza. Le art industries e i nuovi paradigmi di produzione nella new media art contemporanea. Mimesis Edizioni, 2018.
- [5] Manovich, Lev. Il linguaggio dei nuovi media. Edizioni Olivares, 2002.
- [6] McLuhan, Marshall. Gli strumenti del comunicare. Il saggiatore, 2005.
- [7] Michetti, Giovanni. Il paradosso della conservazione digitale: riflessioni sull'autenticità (2010), in <http://digitalia.sbn.it/article/view/237>
- [8] Muñoz Viñas, Salvador. Teoria contemporanea del restauro. Castelvecchi, 2017.
- [9] Purgar, Kresimir. Iconologia e cultura visuale. W. J. T. Mitchell, storia e metodo dei visual studies. Carocci Editore, 2020.
- [10] Quaranta, Domenico. Media, New Media, Postmedia. Postmedia books, 2010.
- [11] Variable Media Network website: <https://variablemedia.net/e/index.html>

# The challenges of preservation: digitizing graffiti in the urban landscape

Massimiliano Carloni<sup>1</sup>, Geert J. Verhoeven<sup>2</sup>, Jona Schlegel<sup>3</sup>, Stefan Wogrin<sup>4</sup>

<sup>1</sup> Austrian Centre for Digital Humanities and Cultural Heritage, Österreichische Akademie der Wissenschaften, Austria – massimiliano.carloni@oeaw.ac.at

<sup>2</sup> Ludwig Boltzmann Gesellschaft, LBI ArchPro, Austria – geert@projectindigo.eu;

<sup>3</sup> Ludwig Boltzmann Gesellschaft, LBI ArchPro, Austria – jona.schlegel@arch-pro.lbg.ac.at

<sup>4</sup> SprayCity, Austria – stefan.wogrin@spraycity.at

## ABSTRACT

Graffiti have always been a contested form of human expression. Although they have increasingly been accepted as museum objects or study topics, in many cases they are neglected or even actively rejected as human expressions worth documenting and (digitally) preserving. However, some scholarly initiatives have tried to overcome this lack of coverage of the graffiti scene. INGRID, for example, collects images from Germany over a span of forty years, while SprayCity archives photos of graffiti from all over Austria and beyond. The limitations with pre-existing solutions, however, is that they do not have standardized criteria to create and process the digital surrogates of graffiti, especially because they derive the photos from different sources.

Project INDIGO aims to overcome these difficulties by offering a systematic, standardized and extensive coverage of a specific area of the city of Vienna, the Danube Canal, which has been a hotspot for graffiti production since the 1980s. For this purpose, INDIGO has developed a highly reliable and reproducible workflow that ensures color accuracy and automated orthorectification of the pictures, and has combined it with a curated digital preservation strategy that involves different services, such as the digital repository ARCHE and the spatial database application OpenAtlas.

While drawing on the specific experience and knowledge acquired in the course of project INDIGO, this paper aims to offer an overview of the challenges that must be faced when digitizing, documenting and preserving contemporary graffiti. It will focus on three key aspects: (1) heritagization, i.e. the process by which an object made by humans such as a graffiti becomes a valuable sample of cultural heritage; (2) FAIRness, the compliance with the so-called FAIR principles (findability, accessibility, interoperability, reusability) in the development of a digital preservation strategy for graffiti; (3) sustainability, i.e. how we can ensure that the data and services provided can integrate into the existing research environment, in order to safeguard the continuous relevance of the products of the research project.

While graffiti are an edge case in the field of current studies, this paper aims to show that discussing the challenges linked to the digital preservation of graffiti can also help us think more carefully about possible improvements in digital preservation strategies for other, more traditionally accepted, kinds of cultural heritage assets.

## KEYWORDS

Graffiti; Cultural heritage; Digital preservation; FAIR principles; Linked Open Data; Sustainability

## 1. INTRODUCTION

Graffiti are ubiquitous, disputable, multifaceted. They often transcend the boundaries between art and vandalism, graphic and textual, criticism and entertainment. Graffiti are becoming an increasingly relevant subject in both academic research and public discourse [13], as they raise pressing questions about intellectual property, the applicability of stylistic and generic categories, and the commercial exploitation of visual works. Accordingly, some efforts have been made to digitize and document them in a more comprehensive way, in order to allow easy retrieval and critical analysis of these materials. For example, INGRID (*Informationssystem Graffiti in Deutschland*; English: information system for graffiti in Germany) [12] digitally archives photos of graffiti in Germany from collections spanning a period of forty years (from 1983 to the present) and allows users to easily annotate and search these digital records. SprayCity [23] documents graffiti in many different cities, with a special focus on Vienna, by including photos taken by the curators, the graffiti artists themselves and the general public.

However, since these archives are based on material from very different sources, they lack strict guidelines and standardization criteria, especially with regard to the photographic material collected, and offer a necessarily unsystematic (though broad) coverage of their reference areas.

The ongoing project INDIGO (*IN-ventory and DI-seminate G-raffiti along the d-O-naukanal*; running from September 2021 to August 2023) [19] takes a different approach, focusing on a more restricted but very active production context, the Danube Canal in the city of Vienna, and aims to digitize and document the graffiti present on almost 13km of surface in a more standardized, accurate, and comprehensive way. INDIGO is the result of a collaboration between several (mostly Austrian) institutions and is funded by the Heritage Science Austria program of the Austrian Academy of Sciences (ÖAW). The Danube Canal has been the hotspot of graffiti production in Vienna since the 1980s and has since attracted a wealth of Austrian and international graffiti artists over the years. More recently, it has also been the site of numerous graffiti-related initiatives, such as tours and workshops; in addition, the growing fascination with its graffitied walls has prompted the opening of several new venues that target an affluent audience, reflecting the ongoing interaction between graffiti and the interests and priorities of different players in the scene.

The aim of this paper is to illustrate some of the challenges faced in the process of digitally documenting and preserving graffiti, as well as the multiple possibilities offered by the use of digital methods in graffiti research. However, while this paper is mainly based on the experience and *interim* results of the INDIGO project, it also aims to offer a broader methodological reflection, applicable not only to graffiti research, but also to other fields and cultural objects.

## 2. HERITAGIZATION

Traditionally, the idea of cultural heritage has been limited to physical objects received from the past, whose artistic merit was already sanctioned by some higher authority and whose preservation was restricted to institutions such as museums and archives. The expansion of scholarly interests as well as technological developments, especially in the last two decades, have led to increased consideration of other forms of human expression, which include more volatile or intangible elements such as oral traditions, music, and films [17, 18].

Grffiti occupy a special place in the realm of cultural heritage [10]. While graffiti are being increasingly recognized as a form of cultural heritage worthy of the same consideration as other objects, they are often not only neglected, but even actively rejected as human expressions worthy of documentation and preservation. For example, searching for ‘graffiti’ and ‘cultural heritage’ in databases of scholarly publications such as Google Scholar (as of 2023-05-05) yields several articles (e.g., [8]) that highlight, more than the closeness between graffiti and cultural heritage, the *damage* made by graffiti to what is usually considered cultural heritage. The very category of street art [3] seems to have caught on precisely to avoid this conundrum: how to give value to something that is contested, often unauthorized, and unconventional by its very nature? How to make a distinction between ‘good’ and ‘bad’ graffiti?

We could use different criteria to determine whether a graffito is worth preserving. For example, we could decide that only graffiti with some artistic value should be selected. However, defining what constitutes art is probably even more subjective and complicated than the definition of cultural heritage itself. Or we could select graffiti that have historical or political relevance. But this is a limitation that does not really apply to cultural heritage, which – on the contrary – is a category that aims to encompass as many cultural aspects of a civilization as possible.

It is difficult – and perhaps impossible – to establish criteria that are clearly defined and unconditionally valid in every situation. The discussion may have to be left open – and may never be solved. Nevertheless, we could still draw three important lessons from the discussion of whether contemporary graffiti should be considered cultural heritage.

First, what determines whether a graffito is preserved is usually a combination of factors and historically contingent phenomena. We may need to look at the overall context in which some graffiti are placed to really understand their cultural significance. While a single graffito on any wall in Vienna might appear to be the irrelevant result of an extemporaneous act of vandalism, the interplay of several thousand graffiti on the entire Danube Canal, the key role that this part of the city has been playing in more recent social developments, the central position that it has acquired in the representation of Vienna itself (for example in advertising, as seen in [5]), all contribute to the Danube Canal’s cultural importance.

Second, time plays an important role. This may seem like an obvious observation. However, time often has a significance that we are not fully aware of when we determine the objects of our academic fields. For example, we may include something in archaeology because we consider it old enough to be an object of some value and worthy of study. However, the boundaries of what we consider ‘archaeological’ are constantly shifting. What in the 1920s was considered to be secondary material that could be discarded is now considered to be a precious testimony to an age long since lost [6]. But time can only play its game if a process of preservation takes place, either intentionally (as in libraries or archives) or accidentally (as in the case of Pompeii).

This leads to a third observation. Preservation is often a major factor in *creating* cultural heritage, or in the words of Laurajane Smith: “heritage is heritage *because* it is subjected to the management and preservation/conservation process, not because it simply ‘*is*’” [15, p. 3]. The act of preservation itself has an effect, either directly or indirectly, on the material

being preserved. Preservation is a process that, given the assumptions that shape the scholarly community and the society at large, not only aims to maintain the materials in the state in which we received them, but also legitimizes those materials and gives them an institutional/communal framework in which to be represented. We should think of digital preservation – and any form of preservation, for that matter – as more active (and sometimes even more invasive) than we usually do.

### 3. FAIRNESS

Digitizing graffiti can be a daunting task. In some cases, graffiti may be placed in difficult locations, such that it is only possible to take photos from angles that give a partial view of the artifact. In addition, since they are in most cases located in open air, one must always consider the influence of external light and possible shadows on the color rendition of the photo. Project INDIGO has developed a workflow that enables accurate color rendition [11] and automatic production of orthophotos [21] at predefined spatial resolutions [19]. The standards held by project INDIGO in the digitization process allow direct comparison between the digital representations of different artifacts and enhance the reusability of these resources, as recommended by the FAIR principles [22]. In fact, standardization operates at different stages of the workflow, both in the digitization phase, where specific conditions must be met to allow for comparable results, and in the post-processing and curation phase, where the acquired data must be brought into compliance with explicitly stated criteria and guidelines.

Especially in the case of graffiti – but also for any other kind of material – the process of standardization should begin long before the final curation of data for preservation purposes. While the outcome of many research projects is a valuable effort and a trove of further possibilities, it is often the workflow itself that can ensure that such data are reusable and valuable to other scholars. Moreover, strict guidelines and good documentation are essential because the digital surrogate is often all we have of a graffiti. A digitized graffiti is not only a backup of the physical object or a more comfortable, alternative way of analyzing the object. In many cases, it is the only way to study the object. To perform automated and reliable analyses, it is necessary that all images are acquired and pre-processed according to stable and explicitly documented criteria. When created and annotated according to clear guidelines, a graffiti dataset opens up a world of possibilities, like training a machine learning model.

Findability and accessibility also play a key role in graffiti research. Since graffiti are ephemeral and often on the fringe of academic studies, very few catalogs or other forms of collections are available. Therefore, it is important that graffiti can be identified and retrieved quickly and reliably. And actually, one of the advantages that make graffiti particularly suitable to modern semantic technologies is that they are, by their very nature, ‘linked’.

Graffiti are contextual and interconnected at several levels. For example, there is a strong connection between a graffiti and its physical location: pre-existing elements of the urban setting can be integrated into the graffiti itself, or specific locations can provide value or an additional meaning to a graffiti. Graffiti can also interconnect according to a sequential arrangement, because a graffiti might refer (in a more or less explicit way) to a previous one; might supplement or complete another creation; or might even overlap with another graffiti or completely cover it, sometimes as an expression of rivalry between two graffitiists (a phenomenon known as ‘crossing’). In this sense, the graffiti landscape creates a sort of contemporary stratigraphy typical of the Anthropocene. Furthermore, graffiti can create links between different parts of a city, different cities or different countries, when creators leave their mark in several places.

All these peculiarities require that the relationships between graffiti, as well as those between graffiti and other kinds of objects or entities, are described in the most detailed and meaningful way. Linked Open Data thus establishes itself as the best conceptual framework to represent this information and link it with other sources. In the context of project INDIGO, technical, administrative and descriptive metadata are stored both in the image files themselves (as embedded Exif or IPTC) as well as in XMP carrier files [1], where they are modeled according to an RDF-based ontology. Information regarding the authorship, physical aspects, and themes of the graffiti is recorded according to the CIDOC CRM ontology [4] for cultural heritage. Concepts and categories like style elements are referred to by newly minted URIs, for example in a SKOS-based thesaurus stored on a stable server [14]; where possible, a match is established with existing concepts, for example from the Getty Art & Architecture Thesaurus [7]. Other types of entities, like artists, are referred to by means of existing Wikidata identifiers, or new Wikidata identifiers are contributed if not already present. The ubiquity of RDF-based standards on the whole scale of the digitization and annotation process thus facilitates data exploration and querying and contributes to the FAIRness of the resulting dataset.

### 4. SUSTAINABILITY

With graffiti, exhaustiveness is almost impossible. At any moment of the day, a new graffiti may be created; graffiti may be located in places that are difficult to access and survey; sometimes they may be removed by public authorities or private

initiatives. The end date of an academic project is only an arbitrary (though necessary) point in time, because the corpus of graffiti, even in a limited area, constantly acquires new items. From this point of view, we have to deal with a situation that is very different from that of more closed (or slower growing) corpora.

Even when we can treat a set of graffiti (chronologically and/or geographically determined) as a somewhat closed corpus, the diversity of approaches, interpretations, and even terminological uses present in the community demands that a workflow aimed at describing and preserving graffiti be as robust as it is flexible. It must be capable of giving space to alternative views on the same artifacts or concepts, and it must be able to accommodate updates and constant improvements without too many problems, as new elements come to the surface that may shed light on the existing material.

Therefore, when it comes to sustainability in relation to graffiti, the concept takes on additional connotations and is linked to a series of specific steps that need to be taken at each stage of an academic project. It is not only a matter of keeping a service running as efficiently and economically as possible, but also, for example, of providing a way to integrate it with other services and applications. This will help keep the service (and the data it provides) relevant in the future.

In the case of project INDIGO, the digital repository ARCHE (*A Resource Centre for the HumanitiEs*) [2, 16] represents the backbone from which the other applications used or developed within the project can retrieve data and metadata. ARCHE offers an easy-to-use and thoroughly documented API [24] that enables access to data in several formats and through diverse dissemination services. One of them is based on the popular ExifTool [9] Perl library and command-line tool, and allows to extract metadata of different formats (Exif, IPTC, XMP) embedded in the images. Thanks to this dissemination service, applications like OpenAtlas [20] can retrieve the information needed with a simple API request and create a spatial database of the graffiti.

Since OpenAtlas can also serve as a tool by which curators can semantically annotate graffiti, the data flow does not only go in one direction. It is possible to re-ingest the image files (and their embedded metadata) into the ARCHE repository in a smooth way, and update the metadata used for internal organization of the repository. This way, digital archiving becomes not only a one-time process that occurs at the very end of a project, but a continuous operation that is integral (and beneficial) to the research workflow.

When considering sustainability, it is also important that a graffiti research project establishes a close connection with the community (academic and non-academic) of interest, because it is one of the tasks of such a project to give back to the community on which it primarily relied – for example, by creating interactive presentations of the collected material, by organizing exhibitions or workshops that can appeal to a broad audience, or by stressing the need to preserve graffiti. This is all the more important since the community will be able to keep the project alive if it is interested, either with additional funding, onboarding in other pre-existing initiatives, or more informal engagement.

Thus, a digital repository should not be seen only as a last resort for data, where the data must be archived in order to comply with regulations by funding bodies and be retrieved only in case of emergency. It can – and should – also serve as a starting point for further projects.

## 5. CONCLUSIONS

There are challenges to digitizing and preserving graffiti that go beyond the challenges we face with other cultural heritage assets. However, discussing and finding solutions to the challenges associated with graffiti can help us improve digital preservation strategies for other materials. For example, it may provide digitization standards that allow for more color-accurate acquisition of images of manuscripts (thus facilitating comprehensive analyses of the physical properties of such artifacts). Or it may provide the impetus for action on digital preservation of still undervalued materials.

## 6. ACKNOWLEDGEMENTS

INDIGO is funded by the Heritage Science Austria program of the Austrian Academy of Sciences (ÖAW).

## REFERENCES

- [1] Adobe Inc. «XMP (Extensible Metadata Platform)», accessed May 5, 2023, <https://developer.adobe.com/xmp/docs>.
- [2] ACDH-CH. «ARCHE(A Resource Centre for the HumanitiEs)», accessed May 5, 2023, <https://arche.acdh.oeaw.ac.at>.
- [3] Blanché, Ulrich. «Street Art and Related Terms – Discussion and Working Definition». SEAC 1 (2015): 32-39.
- [4] CIDOC. «Conceptual Reference Model», accessed May 5, 2023, <https://www.cidoc-crm.org>.
- [5] DreiÖsterreich. «FlyingFIX - Deine fixe Verbindung am Donaukanal», accessed May 5, 2023, <https://youtu.be/ZEISFkBqkqY>.
- [6] Frederick, U.K. «Out of Time and Place», in: B. David, & I. J. McNiven(Eds.), *The Oxford Handbook of the Archaeology and Anthropology of Rock Art*. Oxford University Press (2018). <https://doi.org/10.1093/oxfordhb/9780190607357.013.44>.
- [7] Getty Trust. «About the AAT», accessed May 5, 2023, <https://www.getty.edu/research/tools/vocabularies/aat/about.html>.

- [8] Gomes, Vera, Amelia Dionisio, & J. Santiago Pozo-Antonio. «Conservation strategies against graffiti vandalism on Cultural Heritage stones: Protective coatings and cleaning methods». *Progress in Organic Coatings* 113 (2017): 90-109.
- [9] Harvey, Phil. ExifTool. Software. <https://exiftool.org/>
- [10] MacDowall, Lachlan. «In Praise of 70K: Cultural Heritage and Graffiti Style». *Continuum* 20 (2006): 471-488.
- [11] Molada-Tebar, A., & G. J. Verhoeven. «Towards colour-accurate documentation of anonymous expressions», in: G. J. Verhoeven, J. Schlegel, B. Wild, S. Wogrin, & M. Carloni (Eds.), *Document | archive | disseminate graffiti-scapes*. Proceedings of the goINDIGO2022 international graffiti symposium. *Urban Creativity* (2023).
- [12] Papenbrock, Martin, Doris Tophinke, & Gudrun Oevel. «INGRID – Informationssystem Graffiti in Deutschland», accessed May 5, 2023, <https://www.uni-paderborn.de/forschungsprojekte/ingrid>
- [13] Ross, Jeffrey Ian, Peter Bengtsen, John F. Lennon, Susan Phillips, & Jacqueline Z. Wilson. «In search of academic legitimacy: The current state of scholarship on graffiti and street art». *The Social Science Journal*, 54 (2017): 411-419. <https://doi.org/10.1016/j.soscij.2017.08.004>
- [14] Schlegel, J., M. Carloni, G. J. Verhoeven, & S. Wogrin. «Making a mark – Towards a graffiti thesaurus», in: G. J. Verhoeven, J. Schlegel, B. Wild, S. Wogrin, & M. Carloni (Eds.), *Document | archive | disseminate graffiti-scapes*. Proceedings of the goINDIGO2022 international graffiti symposium. *Urban Creativity* (2023).
- [15] Smith, Laurajane. *Uses of Heritage*. Routledge, 2006.
- [16] Trognitz, Martina, & Matej Ďurčo. «One Schema to Rule them All. The Inner Workings of the Digital Archive ARCHE». *Mitteilungen der VÖB* 71 (2018): 217-231.
- [17] UNESCO. *Basic Texts of the 2003 Convention for the Safeguarding of the Intangible Cultural Heritage*. 2022 Edition (CLT-2022/WS/3; CLD-1001.22).
- [18] Vecco, M. «A definition of cultural heritage: From the tangible to the intangible». *Journal of Cultural Heritage* 11 (2010): 321-324. <https://doi.org/10.1016/j.culher.2010.01.006>
- [19] Verhoeven, G. J., B. Wild, J. Schlegel, et al. «Project INDIGO – document, disseminate & analyse a graffiti-scape». *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLVI-2 (2022): 513-520. <https://doi.org/10.5194/isprs-archives-XLVI-2-W1-2022-513-2022>
- [20] Watzinger, Alexander, et al. OpenAtlas. Software code. <https://github.com/acdh-oeaw/OpenAtlas>
- [21] Wild, B., G. J. Verhoeven, M. Wieser, et al. «AUTOGRAF—Automated Orthorectification of GRAFFiti Photos». *Heritage* 5 (2022): 2987-3009. <https://doi.org/10.3390/heritage5040155>
- [22] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. «The FAIR Guiding Principles for scientific data management and stewardship». *Scientific Data* 3 (2016). <https://doi.org/10.1038/sdata.2016.18>
- [23] Wogrin, Stefan. «SprayCity.at», accessed May 5, 2023, <https://spraycity.at>
- [24] Żóltak, Mateusz. «ARCHE Suite documentation», accessed May 5, 2023, <https://acdh-oeaw.github.io/ arche-docs>

# The Lions of Venice Arsenal: a neuroscientific approach to Cultural Heritage

Elisa Corrò<sup>1</sup>, Nevio Danelon<sup>2</sup>, Maurizio Forte<sup>3</sup>, Federico Boschetti<sup>4</sup>, Alex Kafkas<sup>5</sup>, Luca Ticini<sup>6</sup>,  
Federico Bernardini<sup>7</sup>, Franz Fischer<sup>8</sup>, Paola Peratello<sup>9</sup>, Francesca Dolcetti<sup>10</sup>, Eleonora Delpozzi<sup>11</sup>

<sup>1</sup> Venice Centre for Digital and Public Humanities, Ca' Foscari University of Venice, Italy – elisa.corro@unive.it

<sup>2</sup> Duke University, U.S.A / Ca' Foscari University of Venice, Italy – neviodanelon@gmail.com

<sup>3</sup> Department of Classical Studies, Duke University, U.S.A – maurizio.forte@duke.edu

<sup>4</sup> Venice Centre for Digital and Public Humanities, Ca' Foscari University of Venice, Italy – federico.boschetti@unive.it

<sup>5</sup> Faculty of Biology, Medicine and Health, University of Manchester, United Kingdom – alexandros.kafkas@manchester.ac.uk

<sup>6</sup> Department of Psychology, Webster Vienna Private University, Vienna – luca.ticini@webster.ac.at

<sup>7</sup> Venice Centre for Digital and Public Humanities, Ca' Foscari University of Venice, Italy. – federico.bernardini@unive.it

<sup>8</sup> Venice Centre for Digital and Public Humanities, Ca' Foscari University of Venice, Italy – franz.fischer@unive.it

<sup>9</sup> Ca' Foscari University of Venice, Italy – paola.peratello@unive.it

<sup>10</sup> University of Essex, United Kingdom – francesca.dolcetti@gmail.com

<sup>11</sup> Ca' Foscari University of Venice, Italy – eleonora.delpozzi@unive.it

## ABSTRACT

In this contribution, the authors discuss the preliminary outcomes of a neuroscience research project conducted during the Summer School in Digital and Public Humanities 2022 organized by the Venice Centre for Digital and Public Humanities (VeDPH). The research was based on a series of experiments combining electroencephalography (EEG) and eye tracking to investigate the perception of Cultural Heritage (virtual and real), in this case, the statues of lions in the Venice Arsenal. Particular attention was given to the Piraeus Lion, today at the main entrance to the Arsenal, noteworthy for having been inscribed in the second half of the 11th century AD by Varangian mercenaries with three runic inscriptions, today heavily eroded due to weathering and air pollution. This lion was originally located in the port of Piraeus (Athens) and dates back to the 4th century BC. It was stolen as war booty by a famous Venetian naval commander, Francesco Morosini in 1687 during the Ottoman Venetian war. Among neuroimaging techniques, EEG is used to detect specific states of mind, such as mental alertness, cognitive distraction and learning ability in normal and virtual reality conditions. Specifically, in cultural sectors eye tracking technology is useful to investigate actual involvement in sites of cultural interest.

The data analysis showed changes in the perception of the artifact among different audiences, i.e. women and men, residents and tourists, people with and without previous knowledge on the artifact's historical background. This entails a new method of interactions between a single person, the artifact and the environment. Indeed, this experiment may indicate diversified and accelerated learning and processing, a deep immersiveness of information and different levels of feedback.

## KEYWORDS

Heritage; Neuroscience; EEG; Eye-tracking; Runic inscriptions; Venice Arsenal

## 1. INTRODUCTION

The use of digital technologies can promote critical thinking, creating new experiences to transfer and facilitate student's and citizen's learning. This is very important for future construction of institutional and professional identities, and on visitors' experiences within a possible change in museum practice (Corrò 2021).

Digital technologies are essential for public engagement. The use of immersive and Brain Computer Interface (BCI) technologies could be a means for (global) communication of memories, feelings and sensations during, for instance, site-specific visits, becoming very useful techniques for research and education, not only for dissemination. Advances in neuroscience allow us to conduct quantitative analysis on complex cognitive processes of primary emotions and in specific contexts of observation in the arts and literature (Pearlman 2017). Among other neuroimaging techniques, electroencephalography (EEG), that is the recording of the electrical activity of the brain, allows for the detection of a range of brain waves (alpha, beta, theta and delta) that are characteristic of different rational or emotional mental states, such as interest or excitement psychic states, such as psycho-physical and pathological. In the last decades, BCI technology has been used to explore the frame of mind in several disciplines. In particular, brain waves are being studied to analyze the emotional state, creating data-patterns as a response to visual images that lead to the identification of, for instance, emotions in children with special educational needs (Lekova et al. 2018). EEG is also used to detect specific states of mind,



such as mental alertness, cognitive distraction and learning ability in normal and virtual reality conditions. Moreover, in cultural sectors EEG is used to investigate actual involvement in sites of cultural interest as well as in artistic expressions, such as dance and somatic practice (gestural dialogue), theater, music and performance (Hieda 2017; Paradiso 1999). EEG and eye-tracking are all tools that represent the intersection between different research topics, such as archaeology, cognitive science and neuroscience. For instance, the use of eye tracking allows appreciation of the multimodal nature of vision. This involves the activation of brain, sensory, and emotion-based circuits that can interface with the EEG. Indeed this kind of approach led to reconsider the concept of virtual archeology in the light of the new and stimulating progress of digital archeology, with those relationships and interactions between users and virtual ecosystems, or an environment in which the the observer (the researcher) and the observed (the artifact, the context) are within the same system (Forte, White, Straneva and Woytowicz 2022; Forte et al. 2022).

The ‘Lions of Venice Arsenal’ project aims to investigate experiences and deep interactions that people have with artworks, experiencing the Cultural Heritage (virtual and real), with the use of a combination of EEG and eye tracking methodologies. In particular, the project started from the Venetian Heritage focusing on the symbolism of the lion.

The lion has been the emblem of Venice since the Middle Age as the symbol of Saint Mark, patron of the city, as it appears on the flag of the Republic. Through the centuries, people focused all their efforts on collecting lion sculptures from all over the Mediterranean: from the one of the Piraeus stolen during the war with Ottoman Empire in the 16th century, to the winged one in St. Mark’s square, located on a high column as a symbol of the Saint himself and at the same time of the Venetian State. On the other hand, it seems that Venetians do not identify the lions of the Arsenal with the Lion of Saint Mark. They generally consider them as spoils of war. From this perspective, one of the aims of the project is to clarify this aspect by analyzing this attitude in a more systematic way.

The research started from the analysis of the Piraeus lion [L1], originally located in Athens and dating back to the 4th century BC. It was stolen as war booty by a famous Venetian naval commander, Francesco Morosini in 1687 during the Ottoman Venetian war. On his return to Venice, he set the Piraeus lion at the main entrance to the Arsenale, where it still stands today. The white marble statue is depicted in a sitting pose, it is 3 meters high, and it is particularly noteworthy for having been inscribed in the second half of the 11th century AD by Varangian mercenaries with three runic inscriptions. The graffiti is heavily eroded due to weathering and air pollution in Venice, making many of the individual runes barely legible (Snaedal 2016). From this perspective, the comparison with the other three Arsenal lions seems to go in the same direction: the Hephaisteion lion [L2], originally located in Athens, dated back to the 4th century BC (Sacconi 1991), the Delos lion [L3], dating back to the 6th century BC and, finally, the Molosso lion [L4], are all indeed rich of relevant details, changing or adding new heads as that of a feline, while the original was canine (Giglioli 1952).

The research questions drew on how people perceive the artifact among different audiences, i.e. women and men, and people with and without previous knowledge on the artifact's historical background. Specifically, we already conducted two experiments on different case studies involving both EEG and eye-tracking techniques. Additionally, we will experiment with a VR headset provided with eye-tracking, pupillometry and heart-rate sensors in the final step of the project scheduled for summer 2023 (see below). People involved in the project ranged from art historians to neuroscientists. In the first test (Orient Experience) a VR scenario was created in which the subject could move freely and explore three different urban settings (Mamluk Cairo, Timurid Samarkand ed Angkor) in three different moments of the day (day, sunset and night). The functionality of the Oculus Quest virtual reality (VR) headset in combination with the Emotive Insight (EEG) brain sensor was tested in several sessions. We wanted to verify if the different lighting conditions could to some extent influence the perception of the urban landscape from an emotional point of view. We measured a greater intensity of beta waves (13-30 Hz), characteristics of waking conditions and concentration, for the diurnal scenario and a prevalence of alpha waves (8–12 Hz), characteristics of the state of relaxation before falling asleep, for the nocturnal scenario.

In the second experiment (The Lions of Venice Arsenal) we focused on the eye-tracking technique in an outdoor setting. A group of 10 people, divided by gender, was involved for 40 seconds in a narrative experience while watching the statue of the Piraeus lion. More in detail, we wanted to map the intensity and insistence of the informed observer’s gaze onto the texture of the 3d lion model through the technique of the heatmap (color coded data visualization). The outcomes showed that the most damaged sections of the graffiti was neglected by most of the people.

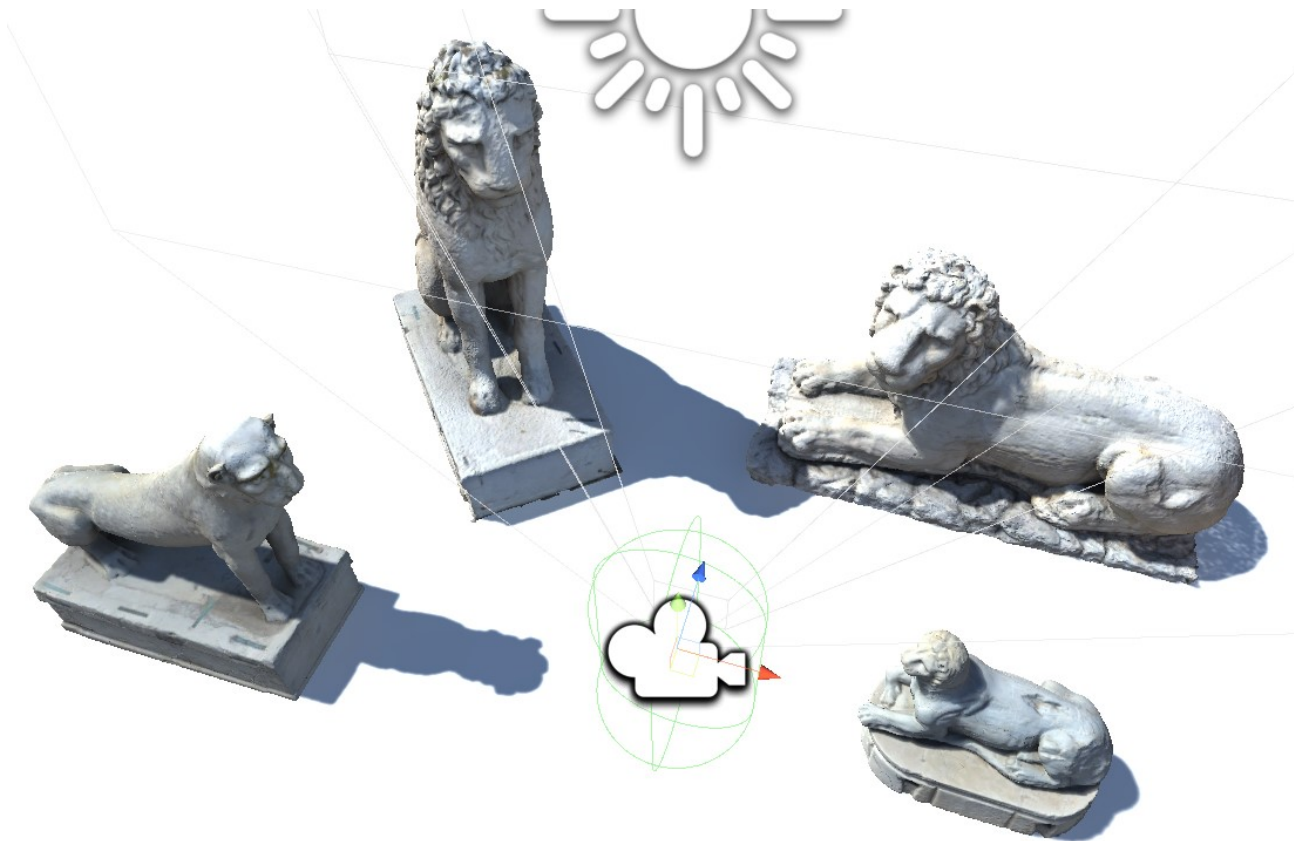


Figure 1. Digital models of Venice Arsenal lions.

A final experiment (Neurocities and Ruinscapes) in collaboration with Duke University<sup>1</sup> will be set in summer 2023 and will involve the participation of a larger group of people in order to focus on the perception of the landscape highlights in relation to Piranesi's views of Rome.

The project is also aimed at promoting a sustainable accessibility to the minor venetian cultural attractions that tend to be overlooked in the richness of the city's heritage. The Arsenal lions became the protagonists, together with other artworks (the Tetrachs of Venice), of the 'Hidden in Plain Sight' virtual exhibition<sup>2</sup>.

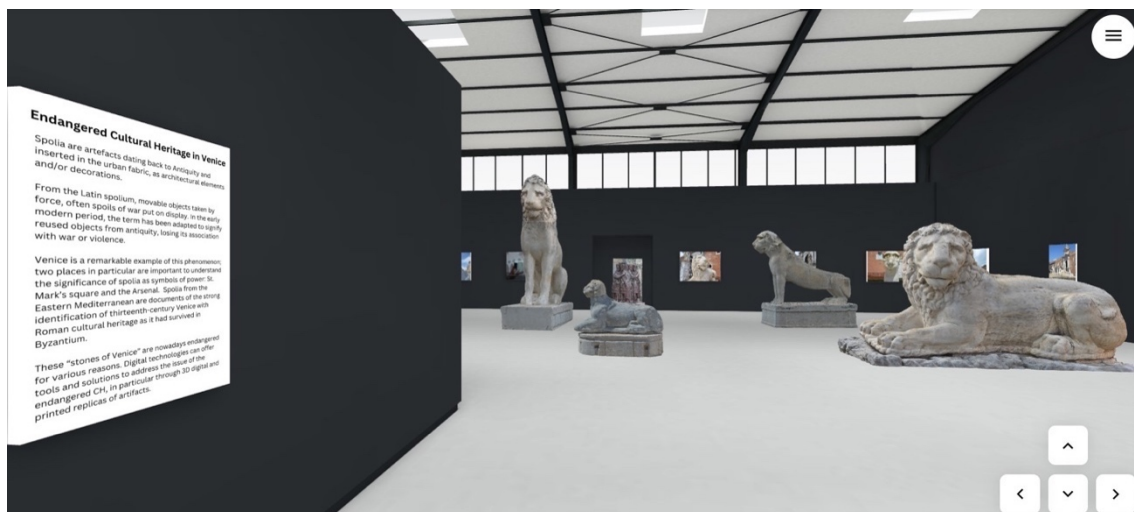


Figure 2. Preview of 'Hidden in Plain Sight' virtual room.

<sup>1</sup><https://bassconnections.duke.edu/project-teams/neurocities-and-ruinscapes-reconstructing-ancient-cities-and-ruins-using-virtual>

<sup>2</sup> 'Hidden in Plain Sight: an (un)curated exhibition on the Tetrachs and Arsenal Lions of Venice' was a Virtual Exhibition curated by Elisa Corrò and Francesca Dolcetti, organized by the Venice Centre for Digital and Public Humanities (VeDPH) in collaboration with the Department of Humanities, the Centro Studi Archeologia Venezia (CeSAV), the DiSpLay Project - Digital Spolia Layering (Ca' Foscari University of Venice, unive.it/display) and the ICPT SciFabLab. The exhibition was online from November 15, 2022 to January 15, 2023.

Within this virtual gallery space we recreated a virtual museum storage room to narrate the story of these artifacts. We chose to display close up pictures as the main subject of the exhibitions, to offer a different perspective and show details of the lions and Tetrarchs that might go unnoticed. Casual shots that symbolize particular meanings and perceptions. An experience that is enhanced by the possibility to interact with the 3D digital replicas in the gallery space. Accompanying these exhibits are various media contents, created using both digital and printed replicas, the lions and Tetrarchs, to discuss the issue of the endangered Venice Cultural Heritage and tell alternative, more playful stories about it.

## 2. ACKNOWLEDGEMENTS

The Lions of the Venice Arsenal Research project - Neuroscience for Cultural Heritage is a project of: Maurizio Forte and Nevio Danelon (Duke University), Federico Boschetti and Elisa Corrò (Ca' Foscari University), Luca Ticini (Webster Vienna Private University), Alexander Kafkas (University of Manchester).

For the virtual exhibition, curated by Elisa Corrò and Francesca Dolcetti, special thanks go to the Department of humanities of Ca' Foscari University, the Centro Studi Archeologia Venezia (CeSAV), the DiSpLay Project - Digital Spolia Layering (Ca' Foscari University of Venice, [unive.it/display](http://unive.it/display): Luigi Sperti, Lorenzo Calvelli, Myriam Pilutti Namer, Eleonora Delpozzi), the ICPT SciFabLab, Federico Bernardini, Eleonora Delpozzi, Nevio Danelon, Paola Peratello and all the participants in the Venice Summer School in Digital and Public Humanities 2022.

## REFERENCES

- [1] Corrò, Elisa. 'Heritage Thinking. Approaching Communities with Communication in the Venice Countryside'. *KnowledgeScape Insights on Public Humanities* 25 (2021): 135-159. <http://doi.org/10.30687/978-88-6969-542-1/007>
- [2] Forte, Maurizio, Ferrara, Vincenza, Mingione, Marco, Alaimo Di Loro, Pierfrancesco, Giorgi, Andrea, Menicocci, Stefano, Babilonia, Fabio, Iosa, Marco, Della Rocca, Carlo, Lecce, Vittoria, Nizzo, Valentino. 'Vedere con la mente. Il progetto NeuroArtifacts'. *Archeologi&. Storia, Antropologia, Museologia, Arte* 67 (2022): 18-29.
- [3] Forte, Maurizio, White, Leonard, Straneva, Kate, Woytowitz, Sean. 'Virtual reality and neuroarchaeology: visual perception and cognition of an archaeological excavation'. Banzi, A (ed.), *The Brain-Friendly Museum: using Psychology and Neuroscience to Improve the Visitor Experience* (2022). <https://doi.org/10.4324/9781003304531>
- [4] Giglioli, G. Quirino. 'I leoni dell'Arsenale di Venezia'. *ArchCI IV* (1952): 1-9.
- [5] Hieda, N. 'Mobile Brain-Computer Interface for Dance and Somatic Practice'. *UIST '17: Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology = Conference Proceedings (Québec, Canada, 22-25 October 2007)*, (2017): 25-6. <https://doi.org/10.1145/3131785.3131803>
- [6] Lekova, Anna, Dimitrova, Maya, Kostova, Snezhanka, Bouattane, Omar, Ozaeta, Leire. (eds). 'BCI for Assessing the Emotional and Cognitive Skills of Children with Special Educational Needs'. *IEEE 5th International Congress on Information Science and Technology (CiSt), Marrakech, Morocco* (2018): 400-403. <https://doi.org/10.1109/CIST.2018.8596571>
- [7] Paradiso, Joe. 'The Brain Opera Technology: New Instruments and Gestural Sensors for Musical Interaction and Performance'. *Journal of Music Research* 28 (1999): 130-49.
- [8] Pearlman, Ellen. 'Brain Opera. Exploring Surveillance in 360-Degree Immersive Theatre'. *PAJ: A Journal of Performance and Art* 39-2 (2017): 79-85.
- [9] Sacconi, Antonella. 'L'avventura archeologica di Francesco Morosini ad Atene (1687-1688)'. *Rivista di Archeologia* 7 (1991).
- [10] Snaedal, Thorgunn. 'Runes from Byzantium: reconsidering the Piraeus Lion'. *Byzantium and the Viking World* (2016): 187-244.

# L'immaginazione tra 0 e 1.

## Alcune considerazioni su memoria e oblio nell'era digitale

Ivana Brigida D'Avanzo<sup>1</sup>

<sup>1</sup> Università degli Studi di Perugia, Italia – ivanabdavanzo@gmail.com

### ABSTRACT

Il presente contributo intende rileggere la dinamica memoria-oblio alla luce delle nuove scoperte tecnologiche e della comunicazione digitale. La natura ancipite della memoria umana, ricettacolo di ricordi ma soprattutto attività di relazione – costruzione, *ri-costruzione*, impiego, condivisione – con e dei ricordi stessi è qui considerata in un triplice nesso di *attività* che vede, al polo opposto la “dimenticanza attiva” di matrice nietzschiana e, al centro – in uno spazio di comunicazione non solo possibile ma umanamente imprescindibile – la facoltà immaginativa.

### PAROLE CHIAVE

Memoria; oblio; immaginazione

### 1.

Venga assunta come punto d'avvio di questa indagine una definizione di memoria, volutamente semplice – non per questo semplicistica – che ponga immediatamente all'attenzione del lettore la sua connaturata duplicità: con il termine memoria, infatti, si intende, da una parte, la facoltà di molti organismi – ognuno secondo diverse e specifiche modalità – di tenere traccia delle proprie esperienze le quali, registrate in un dato passato, sono rese ancora fruibili nel futuro; ma anche, parallelamente, lo spazio nel quale tali tracce sono conservate. La memoria è dunque ricettacolo di ricordi ma anche attività di relazione – costruzione, *ri-costruzione*, impiego, condivisione – con e dei ricordi stessi; attività, quest'ultima, che non può prescindere – come sarà ribadito con maggiore enfasi a breve – dalla cooperazione, come in una «funzionale mescolanza» ([1]: 43), con la facoltà immaginativa. Ogni discorso filosofico sulla memoria non può prescindere dal richiamo ad ambedue le suddette polarità, che si intersecano e si implicano vicendevolmente.

Lo spazio di archiviazione di dati di cui ogni sistema informatico è dotato viene definito, per «analogia semantica» ([10]: 28-29), *memoria*. Non vi è dubbio, infatti, che sul piano digitale esista una corrispondenza con la capacità umana di *salvare* dati; naturalmente, di questa corrispondenza occorre, seppur brevemente, mostrare il discrimine, tanto dal punto di vista *quantitativo* che *qualitativo*. Anche Luciano Floridi analizza i problemi relativi alla memoria digitale considerando la polarità quantità-qualità, e la reciproca influenza che l'un polo ha sull'altro<sup>1</sup>. Nella fattispecie, lo studioso italiano lascia emergere l'intreccio problematico che scaturisce dalla costante produzione di dati – «dal 2007 il mondo ha prodotto più dati di quanti ne possa immagazzinare» ([6]: 21) – e dall'incapacità di gestire i big data, di *filtrarli*, vale a dire, concretamente, di discernere e quindi di discriminare la singola informazione ricercata dalle altre *incalcolabili* informazioni a disposizione dell'utente: il proverbiale ago nel pagliaio, spesso ma non volentieri. «Metà dei nostri dati sono spazzatura – afferma Floridi con un certo richiamo al *Memorioso* borghesiano<sup>2</sup> – il fatto è che non sappiamo quale metà» ([6]: 22). E allora, cosa buttare e cosa salvare? La necessità, nel corso degli ultimi decenni, di coniare nuove e sempre maggiori unità di misura di byte in espansione continua illudono di *poter salvare pressoché tutto*. Ma, a dispetto di una più superficiale credenza, «le ICT posseggono un genere di memoria che dimentica» ([6]: 19): il problema, naturalmente, non è tanto «il nostro smartphone troppo pieno perché troppo carico di foto» bensì la proiezione del problema «su scala globale». Se il problema dello storico è sempre stato “cosa salvare”, ora, «nell'iperstoria, salvare è opzione di default. Il problema diventa cosa cancellare» ([6]: 22).

<sup>1</sup> Come già *La scienza della logica* hegeliana ha mostrato: «la variazione del quanto è anche un mutamento di qualità» ([7]: 301-302).

<sup>2</sup> «La mia memoria, signore, è come un deposito di rifiuti» ([3]: 712).

## 2.

Dimenticare non è una semplice *vis inertiae*, come ritengono i superficiali, ma piuttosto una facoltà attiva, positiva [...].

Chiudere di tanto in tanto porte e finestre della coscienza; restare indisturbati dal rumore e dalla lotta con cui il mondo sottostante degli organi posti al nostro servizio svolge la sua collaborazione ed opposizione; un po' di silenzio, un po' di *tabula rasa* della coscienza, *affinché vi sia ancora posto per il nuovo* [;] è questo il vantaggio – come si è detto – della dimenticanza attiva, una guardiana, per così dire, una sorvegliante dell'ordine spirituale ([11]: 256. Corsivo mio).

L'elogio nietzschiano dell'oblio e del beneficio *vitale* della capacità di dimenticare – della dimenticanza attiva – è questione filosoficamente ben nota. Non solo nella *Seconda inattuale* sulla storia ma anche e forse ancora più esplicitamente qui, ad *incipit* della Seconda dissertazione della *Genealogia della morale*, Nietzsche avverte dell'impossibilità di ricercare felicità, serenità, speranza senza aver sviluppato una prioritaria quanto necessaria «capacità di dimenticare» ([11]: 256).

Ora, non si tratta qui soltanto di riconoscere la necessità dell'oblio, finalizzata al sano funzionamento della memoria, la quale, infine, *serve alla vita*. Va da sé che la costruzione della nostra propria identità personale non può darsi senza una dinamica relazione con la memoria, vale a dire con la dialettica memoria-oblio: noi siamo ciò che ricordiamo tanto quanto ciò che dimentichiamo. E non essendo oggi sensato – né tantomeno possibile, forse – scindere nettamente una dimensione *solamente* umana della nostra memoria da una controparte digitale, è chiaro che occorre tenere presente quanto e in che misura la memoria digitale – in ogni sua forma – collabora nella costruzione della nostra identità. Ma una cosa è l'abitudine, ormai irrinunciabile, di affidare i nostri ricordi, o almeno una parte di essi, a supporti esterni – sebbene sempre a portata di smartphone –, distribuendo *i nostri dati* tra l'archivio del corpo o quello delle macchine; ben altra cosa, invece, è la delega – ammessa la possibilità di realizzarla – della più umana facoltà che corrisponde alla dialettica triadica memoria-immaginazione-oblio.

Assunta, pertanto, come presupposto biologico e cognitivo, la *funzione attiva* dell'oblio<sup>3</sup>, l'intento è quello di andare ad indagare proprio nell'incrocio tra i luoghi della memoria e quelli dell'oblio uno spazio di relazione tra due dimensioni solo apparentemente incomunicabili. È proprio qui, infatti, in questa fertillissima – dal punto di vista della libertà creativa – terra di mezzo, che si gioca uno scarto fondamentale con quella che si può definire la più classica struttura della memoria digitale, ovvero quella che si basa sullo *storage and recovery*. Archiviare e recuperare – o meglio, archiviare al fine di recuperare – è, in effetti, il compito che ci aspettiamo una memoria assolvere: sia essa quella dei nostri dispositivi personali, laddove tanto lo *storage* quanto il momento di *recovery* sono a nostra propria discrezione e responsabilità, sia invece quella dei motori di ricerca e delle banche dati che, debitamente caricate, sono infine messe a disposizione dell'utente. Presupposto ed obiettivo di questo paradigma storage-recovery è che l'informazione archiviata nel passato (quand'anche estremamente prossimo), sia recuperata nel futuro (persino un immediato futuro<sup>4</sup>) come totalmente integra, senza alcuna perdita di dati. Si sa che le memorie digitali, in massima parte, sono in grado di restituire *ripetutamente* un dato identico a sé stesso; che, di contro alla fallibilità della memoria umana per la quale vale la proporzione *perdita di tempo-perdita di dati*, sono in grado di vincere questa sfida contro il tempo. Ma in merito a questa presunta vittoria vorrei fare due considerazioni.

La prima concerne possibili falle nel paradigma digitale sopra descritto, meno circa il momento d'archiviazione quanto più rispetto a quello della restituzione. Basti pensare alle tantissime pagine abbandonate su Internet o ai diffusi casi di link scaduti: minimi ma efficaci esempi di quello che Floridi ha definito «paradosso della preistoria digitale»:

le ICT non stanno conservando il passato per metterlo a disposizione del futuro dal momento che ci fanno vivere in un eterno presente [...]. Una pagina web che si aggiorna costantemente è un sito che non conserva memoria del proprio passato, e lo stesso sistema dinamico che consente di riscrivere migliaia di volte lo stesso documento rende altamente improbabile la conservazione delle versioni precedenti per un esame futuro<sup>5</sup>.

Persino la “aspettativa di vita” dei sistemi informatici (MTBF, *Mean time before failure*) non è così lunga come si possa ritenere. E non v'è dubbio che le informazioni possano essere spostate su nuovi supporti o cloud e immagazzinamenti

<sup>3</sup> Si vedano in proposito le considerazioni di Nicola Russo, che, classificando l'oblio come «facoltà essenziale sia dal punto di vista biologico, che da quello cognitivo», contestualizza i nessi traccia-simbolo e memoria-oblio, con le relative implicazioni digitali, sul «piano elementare dell'ontologia», vale a dire considerando un terzo fondamentale nesso: «vita e conoscenza» ([15]: 76).

<sup>4</sup> Riccardo De Biase, a proposito del dominio della velocità in campo digitale, ha parlato di una vera e propria «tachicrazia» ([5]: 191).

<sup>5</sup> [6]: 19. Le implicazioni della rivoluzione digitale sul piano della temporalità sono state oggetto di numerose indagini. Mi limito qui a ricordare la proposta di Castells di un *tempo senza tempo* ([4]: 491-533). Sempre utile tenere presenti le riflessioni sul tempo nella tarda-modernità del sociologo tedesco Niklas Luhmann [8]. Si veda, su questi temi, pure l'interessante e ricca (anche dal punto di vista bibliografico) analisi di Felice Masi sul tempo “passato prossimo” [9].

online (Dropbox o Google Drive, per portare solo qualche esempio) ma – c'è da domandarsi – tutti i dati sopravvivono a queste transizioni? E, soprattutto, è possibile praticare questa strategia a livello globale?<sup>6</sup>

La seconda considerazione riguarda la suddetta fallibilità della memoria umana di ricalcare un modello digitale capace di una restituzione integrale del contenuto salvato. Ebbene, questa presunta fallibilità, rappresenta piuttosto una prerogativa umanissima, sovrintendente la dimensione creativa, *poietica*, artistica, indispensabile per la costruzione di ogni identità personale.

Fino a poco tempo fa, si riteneva ottimisticamente che le ICT rendessero gli individui più capaci di dare forma alle loro identità personali. Il futuro sembra più sfumato. I ricordi registrati tendono a congelarsi e rafforzano la natura del soggetto. Quanto più accumuliamo ed esternalizziamo i nostri ricordi, tanto più delimitiamo dal punto di vista narrativo la costruzione e lo sviluppo delle nostre identità personali. Accrescendo i nostri ricordi, riduciamo anche lo spazio di libertà di cui godiamo nel ridefinire noi stessi. Dimenticare è parte integrante del processo di costruzione di sé. Una soluzione possibile per le generazioni a venire può consistere nell'essere più parsimoniosi con tutto ciò che tende a cristallizzare la natura del sé ([6]: 82).

Qui – a partire da queste considerazioni di Floridi – torna in gioco la dimenticanza attiva nietzschiana, come condizione indispensabile affinché possa darsi *il nuovo*. Se per le memorie digitali vale la massima del *tertium non datur* – salvare o cancellare, accesso o spento, 1 o 0 – per la memoria umana, tra l'oblio come totale amnesia, come nulla, e la memoria come salvataggio perfetto e senza residui, esiste un terzo polo, centrale e fecondo. Questo è il regno dell'immaginazione – nel senso etimologico di facoltà di produzione di immagini – che non consente mai e in alcun modo una seconda restituzione di ricordi identici a sé stessi. Non esiste memoria senza immaginazione<sup>7</sup>: di fronte ad ogni fenomeno mnemonico – che si tratti di un ricordo quanto mai nitido o di un frammento sfocato – l'uomo si pone in relazione ad esso attraverso la sua immaginazione; ricostruisce, trasforma poco o tanto, consapevolmente o inconsapevolmente, ri-vivifica. Ed è esattamente la suddetta fallacia della memoria umana a consentire questi interventi che generano metamorfosi di immagini mai uguale a sé stesse, mai ri-producibili due volte<sup>8</sup>.

Ricordare è sempre anche ri-creare, *mimesis* attiva<sup>9</sup>. Il filosofo francese Gaston Bachelard, nelle sue riflessioni sull'immaginazione come *rêverie*, ha parlato di una memoria *cosmica*, che restituisce una altrettanto *cosmica* infanzia: ciascuno di noi non possiederebbe così soltanto la propria infanzia, i soli ricordi personali, scanditi da una precisa ed accertata cronologia, bensì un'infanzia assai più ricca, “analogo di una falsa memoria”, “attività di una *meta-amnesia*” ([2]: 113):

Nella loro primitività psichica, Immaginazione e Memoria appaiono in un complesso indissolubile, difficilmente analizzabile se riferito alla percezione. Il passato ricordato non è semplicemente un passato della percezione, nel momento in cui ci ricordiamo, il passato è percepito come valore di immagine. L'immaginazione colora fin dall'inizio i quadri che le piacerà rivivere. Per risalire agli archivi della memoria, è necessario individuare dei valori al di là dei fatti. Non si analizza la familiarità enumerando le ripetizioni [...]. La Memoria e l'Immaginazione competono per restituirci le immagini che appartengono alla nostra vita ([2]: 110).

### 3.

La memoria, ancipite, è sempre possesso e attività. L'utilizzo dei supporti digitali che mettono a disposizione dell'utente – di tutti – banche dati ricchissime è senz'altro una risorsa preziosa, a dispetto delle difficoltà legate al filtraggio dei big data. Ma una delega della dinamica memoria-oblio – che qui si è voluta declinare come triade memoria-immaginazione-oblio – non sembra essere praticabile, a dispetto di ogni intelligenza artificiale.

Non si è trattato qui di proporre una critica dai danni delle memorie digitali. È inevitabile, certo, riferirsi alla memoria digitale con metodologia comparativa con la memoria umana, soprattutto tenuto conto della natura filosofica – e non

<sup>6</sup> Di grande rilievo, pertanto, le iniziative delle associazioni come la National Digital Stewardship Alliance (NDSA), volte alla conservazione dell'eredità culturale digitale. Si rimanda al sito <https://ndsa.org>

<sup>7</sup> La reciproca implicazione di immaginazione e memoria può essere discussa a partire da riferimenti e contesti disciplinari più disparati. Anche dal solo punto di vista filosofico le testimonianze sono incalcolabili. Si sceglie di ricordare qui soltanto la tesi vichiana secondo la quale la memoria, non solo “raccolge come in un recipiente le percezioni acquisite per mezzo dei sensi” ma è *già* “attività produttrice di immagini la facoltà che i greci chiamavano fantasia e che noi chiamiamo immaginativa” ([16]:114). La medesima reciprocità di memoria e immaginazione è sostenuta pure – questa volta sul piano neuroscientifico – da Alberto Oliviero [12] e [13].

<sup>8</sup> A partire dal tema della ripetibilità dell'informazione digitale, sempre a disposizione dell'utente come uguale a se stessa, potrebbe essere interessante – ma si rimanda ad un altro momento e un altro luogo – una linea di discussione che intersechi la questione del nesso umano troppo umano di immaginazione-memoria con i temi dell'arte contemporanea e della sua riproducibilità tecnica.

<sup>9</sup> Sempre utili, in tal senso, le considerazioni sul concetto di *mimesis* del Ricoeur di *Tempo e racconto* [14: 91-141].

immediatamente tecnico-informatica – di questa proposta; ma tale comparazione non è fine a se stessa. Il richiamo alla declinazione che la dinamica memoria-oblio assume sul piano digitale, è stato qui finalizzato – come ogni possibile discorso filosofico sul digitale *in toto*, per ferma convinzione di chi scrive – ad una sempre più approfondita e consapevole analisi dell'*umano* in ogni suo aspetto o prospettiva possibile. Indagare i domini dell'*umano* considerando ma isolando la sfera digitale da tutte le altre – arte, scienza, linguaggio, religione persino – può presentarsi oggi alla stregua di una *mutilazione culturale*. Si è tentato di considerare, pertanto, la dinamica memoria-oblio non tanto *in rapporto alla* memoria digitale ma *tenuto conto della* memoria digitale, come fattore imprescindibile della facoltà che l'uomo contemporaneo ha di esercitare la memoria e l'oblio. Le memorie digitali giocano senz'altro un ruolo cruciale, oggi, nella quotidianità comune come nella ricerca scientifica, come estremamente vantaggioso strumento di indagine. Eppure, considerata la natura policentrica della memoria umana, restano a tali strumenti preclusi, territori dell'*umano* che faticano ad essere circoscritti all'interno di un sistema binario che non ammette uno spazio di relazione tra lo 0 e l'1.

Almeno per ora.

## BIBLIOGRAFIA

- [1] Bachelard, Gaston. La poetica dello spazio, Dedalo, Bari, 2006.
- [2] Bachelard, Gaston. La poetica della rêverie, Dedalo, Bari, 2015.
- [3] Borges, Jorge Luis. Funes, o della memoria in Id., Tutte le opere, vol. I, Mondadori, Milano, 1996: 707-715.
- [4] Castells, Manuel. La nascita della società in rete, Egea, Milano, 2014.
- [5] De Biase, Riccardo. Comprendere il digitale. Un approccio kantiano, in Carbone, Raffaele (a cura di) Modernità e critica, Città del sole, Napoli, 2021: 171-201.
- [6] Floridi, Luciano. La quarta rivoluzione. Come l'infosfera sta trasformando il mondo, Raffaello Cortina, Milano, 2017.
- [7] Hegel, Georg Wilhelm Friedrich. La scienza della logica, UTET, Torino, 2004.
- [8] Lumann, Niklas. The future cannot begin: temporal structures in modern society, in «Social Research» 43, 1 (1976): 130-152.
- [9] Masi, Felice. Just-in-time. Il passato prossimo tra estenuazione, banalità e spreco, in «EDA. Esempi di architettura», numero monografico: Segni del presente. Filosofia della cultura e culture del digitale, Aracne, Roma, 2019, 46-57.
- [10] Melandri, Enzo. La linea e il circolo, Il mulino, Bologna 1968.
- [11] Nietzsche, Friedrich, Genealogia della morale, in Id., Opere di Friedrich Nietzsche, Adelphi, Milano, 1972, vol. VI, t. II: 211-367.
- [12] Oliverio, Alberto. Immaginazione e memoria. Fantasia e realtà nei processi mentali, Mondadori, Milano, 2013.
- [13] Oliverio, Alberto. Memoria e oblio, Soveria Mannelli, Rubbettino, 2003.
- [14] Ricoeur, Paul. Tempo e racconto, vol. I, Jaca Book, Milano, 2008.
- [15] Russo, Nicola. Traccia e simbolo tra memoria e oblio, in Segni del presente: 74-83.
- [16] Vico, Giambattista, De antiquissima italarum sapientia, Id., Opere filosofiche, Sansoni, Firenze, 1971: 55-131.



# A methodological study on the projects developed by the Cologne Center for Humanities (CCeH)

Mauro De Bari<sup>1</sup>

<sup>1</sup> University of Bari Aldo Moro, Italy – mauro.debari@unba.it

## ABSTRACT

This paper exposes the *modus operandi* in creating digital cultural entities (DCE) by an international academic reality, notably the Cologne Center for eHumanities (CCeH), after spending a semester studying in the centre in question. Firstly, an explanation of the mechanisms and methodologies implemented in the digital creation of cultural resources will be provided. Subsequently, the development and preservation methods of digital entities will be analysed. The CCeH dedicates its activities to academic research projects. The main goal is to support teaching and learning, reflecting on specific areas of study by implementing the methodology and processes that manage technology in the cultural field. Once the intent, the goals and the results are explained, some good practices dedicated to creating new digital cultural artefacts are underlined. It is also considered the impact of awareness that the digital entities created produce, more or less consciously, on the users, who are provided with different levels of digital alphabetisation. In particular, how the final user will have specific importance in the design of projects will be shown: changing the user's destination means carefully pondering the interaction with the digital creation, with particular attention dedicated to the achievable goal.

## KEYWORDS

Digital cultural entities (DCE); Methodologies; Digital Artefacts (DA); Academic research projects; Digital Preservation.

## 1. INTRODUCTION

In recent years we have witnessed renewed attention towards the historical dimension of Digital Humanities (DHs) [1]. However, their inclusive and multicultural nature makes a series of actions and considerations necessary for a univocal classification of such a varied sector difficult.

On the other hand, the vastness and heterogeneity of the DHs maintain an undisputed development potential, which allows for the creation of various projects of various kinds, with an impact of great interest for the definition of cultural heritage. In this regard, a great debate opens up about the sustainability and preservation of the digital entities created to preserve cultural memory [2].

In this scenario, cultural institutions have deliberately worked to transform important cultural assets into potential digital assets. The same institutions, after periodic evaluations, try to maintain the digital resources undamaged using conservation techniques lacking a standard methodology. When necessary, they apply interventions and restorations, depending on the entity and its state at different times of the state of conservation.

However, assuming the speed of Digital Transformation, this process becomes more complex and problematic. Technological development negatively affects the long-term preservation of digital assets. As we know, digital preservation is a topic that has been discussed since 1996; the Task Force on Archiving of Digital Information report is a topical document in this regard [3].

However, nowadays, it is necessary to take a step forward. For this reason, one of the most challenging actions that cultural institutions, education and academics should address is to create a digital culture that promotes awareness in a mutual exchange of legal ownership, knowledge, and responsibility in the Community to then, subsequently, protect and safeguard the cultural heritage, digital and otherwise. This assumption is pioneering considering the beliefs of the last decade: where digitization was “the instrument” to safeguarding and administering the cultural heritage (CH), without considering the preservation of the digital entities created adequately[4].

Nowadays, it seems clear that digital preservation has played an undisputed central role: it should no longer be a corollary of cultural heritage or a functional theme of cultural heritage, considering how much cultural institutions are investing in digital innovation. Nevertheless, there is a critical issue: the possibility of losing information and risking entering the Digital Dark Age, considering that data are replacing documents [5].

We currently live the risk of crashing or losing data, and what is solved now, probably in the future, could not. Thus, an extreme necessity to ponder a strategy to preserve data and information: in terms of provenance [5].

Before considering the role of institutions and public entities in producing digital information that must be readable and intelligible in time (this pressure becomes more stringent than before). The challenge is identifying the entire amount of digital content the potential digital heritage can preserve in a short and long time. Until the methods to achieve this are defined, digital data preservation will be one of the biggest challenges[6].

In this global phenomenon, multidisciplinary inclinations are *conditio sine qua non* to preserve digital entities. It is not easy to design a winning strategy. The Academy, Cultural Institutions and CCIs need to balance their needs. For this reason, this paper exposes some projects designed and developed by the CCeH<sup>1</sup> of the University of Cologne<sup>2</sup>, describing some exemplifying examples that qualify the history of the departmental center, showing some good practices methodologies and processes identified in terms of digital creation, innovation, preservation and end-user for which the digital project is intended.

## 2. A FIRST METHODOLOGICAL APPROACH

In the semester spent at the University of Cologne, in particular at the CCeH<sup>3</sup>, the first-level goal of the study conducted was to identify a *fil rouge* among the different digital prototypes designed and created in the Departments of the University of Cologne; for this reason, it has been necessary to plan a strategy to get a sequence of helpful information, analyzing different projects of different nature. First, a questionnaire was submitted for each project representative staff member interviewed.

The arguments of interest were focused on understanding the long-term preservation planned by the teams involved in managing the projects. Because of this, questions were based on four principal aspects:

- funds obtained and dedicated to the scope;
- methodologies and processes used;
- main goals achieved;
- the policy of preservation of the digital cultural artefacts created.

In this perspective, the questionnaire was addressed to a generic target of professionals<sup>4</sup> to avoid falling into specifics. The following four sections of questions were proposed.

- Section one: “Funds”:
  1. Did you apply for investments?
  2. Who was the principal investor?
  3. Who signed the partnership agreement?
- Section two: “Methodologies and Processes”:
  1. Which kind of theories, methodologies and models of publications have you considered to design the project?
  2. Could they be helpful today? Or are they obsolete?
  3. Can you describe the process used and created for this project, explaining the passages from the creation until the end?
  4. Which kind of process did you create to organize the work?
  5. Who are the partners involved?
- Section three: “main goals achieved”:
  1. Can you briefly describe and code the results achieved in this project?
  2. What kind of digital cultural artefacts have you produced?
  3. Which sort of digital protocol have you used for digitizing the artefacts?
  4. Did you create a process or test an existing one?
  5. Have you made Scientific contributions?

---

<sup>1</sup> Cologne Centre for Humanities, Home page, <https://cceh.uni-koeln.de>, (last consulted: 03/03/2023).

<sup>2</sup> University of Cologne, Home page, <https://portal.uni-koeln.de/en/uoc-home>, (last consulted: 03/03/2023).

<sup>3</sup> Cologne Centre for Humanities, Home page, <https://cceh.uni-koeln.de>, (last consulted: 03/03/2023).

<sup>4</sup> It is fundamental to underline that in this context and in the projects analyzed were involved professionals with a specific and a heterogenous background with the aim to merge different knowledge and know-how to achieve satisfying results.

Section four: “the policy of preservation of the digital cultural artefacts created”:

1. Did you create a repository system? Or did you use an existing one?
2. What are you going to do when the funds end?
3. Have you planned a strategy to preserve digital creations made?

During an introductory debate with the Center team, it was better to expand the case studies beyond the initially planned three projects selected, achieving a more concrete result. Initially, the projects analyzed should be the DiXiT project<sup>5</sup>[7]: a comprehensive teaching and research program tailor-made to meet the different needs emerging due to the pressing digitalization of the humanities<sup>6</sup>; part of Lazarus<sup>7</sup> (tripartite project sub-sections: the Meister-Eckhart Archive<sup>8</sup> [8], the Sanskrit Dictionary<sup>9</sup> [9], Paleography Online (PO)<sup>10</sup>[10] and Colonia 3D<sup>11</sup>[11]). Subsequently, the list of projects hypothesized was expanded, to the above projects were added Haller Net<sup>12</sup>[12], (Re)-Collecting Theater History<sup>13</sup>[13], the

---

<sup>5</sup> The project addresses the hermeneutical questions on literature and historical writing and the technical challenges of digitised research and publishing practices. Specifically, it deals with issues such as the application and improvement of existing data standards for the description and coding of documents, the development and integration of digital tools for indexing and scientific presentation of texts and security and usability to long term of digital editions. In August 2017, making DiXiT official came to an end. A miscellaneous volume was published by Sidestone Press, detailing the research and projects presented at the three majors DiXiT conferences.

<sup>6</sup> The project addresses the hermeneutical questions on literature and historical writing and the technical challenges of digitised research and publishing practices. Specifically, it deals with issues such as the application and improvement of existing data standards for the description and coding of documents, the development and integration of digital tools for indexing and scientific presentation of texts and security and usability to long term of digital editions. In August 2017, making DiXiT official came to an end. A miscellaneous volume was published by Sidestone Press, detailing the research and projects presented at the three majors DiXiT conferences.

<sup>7</sup> The focus of the LAZARUS project is the ideal preparation of three sub-projects aimed at sustainable protection and the permanent provision of digital research. Three university collections, evaluated as exemplary, were selected for the project: the Meister Eckhart archive, the Cologne Sanskrit digital lexicon (CDSL), and Palaeography Online (PO). All three collections have in common that their data is already available in different formats.

<sup>8</sup> The MEA project was born with the ambitious idea of reconstructing the Meister Eckhart archive in a digital collection. The Meister Eckhart Archive was owned by Prof. Dr Josef Koch (1885-1967), founder of the Thomas Institute of the University of Cologne, (the renowned research institute on the history of the Middle Ages). The archive includes files, certificates, and correspondence documenting the creation of the entire Eckhart edition from the 1930s to today. The more than one thousand unpublished documents are currently undergoing digitization, cataloguing, and scientific development under the direction of Prof. Andreas Speer (director of the Thomas Institute). The Meister Eckhart Archive catalogues and metadata, still available as Word tables, are now converted into an XML-based template.

<sup>9</sup> The Cologne Sanskrit Digital Lexicon (CDSL) collects lexical data in the Sanskrit language. It contains information from about fourteen, in some pervasive cases, Sanskrit dictionaries and other lexical resources in digital form. The foundations of the digital collection were created over 16 years ago so that older data is no longer up to date. The individual dictionaries represent separate resources with project-specific markup formats and are partially available through a web application. The visual rendering offered to the user reflects the heterogeneity of the material and the increased structure of the collection. The purpose of preparing this lexical treasure is to standardize the markup of various resources and use the Text Encoding Initiative (TEI). The idea behind the Sanskrit project borrows from the concept of up-conversion (low-quality data are converted into a better resolution aimed at allowing better processing and preservation) and format-shifting. One of the first problems in Sanskrit was data conversion, considering that from the 1980s, no method united Latin and Indian characters (devanāgarī). Therefore, the developers had to find a conversion method without using the XML language that does not yet exist. The most functional and immediate response was the invention of a sui generis “standard” protocol. However, since the nineties, the development and implementation of the XML language have made it possible to correct existing typos in the versions previously used.

<sup>10</sup> Paleography Online is configured as an auxiliary e-learning offer for university education (so far, just over 1000 registered users). At its core, it comprises a series of chronologically and thematically organized exercises on the Paleography of Roman antiquity and the Western Middle Ages, which is complemented by an illustrated glossary and an overview of alphabets and abbreviations considered model. As part of the project, functionality and programming must be improved, and technical prerequisites must be created to further develop the offer of online Paleography. The system is currently based on PHP scripts and outdated JavaScript functions. Complete re-modelling, transfer to a modern database system, and integration into a permanent server infrastructure are required. Professor Georg, one of the project’s creators, decided to create a partnership with Cologne by conveying the idea and the first writing formats in HTML language. Arrived in Cologne and aware of the problems existing in MEA, Professor George members of the centre started from almost static content, investing as much as possible in the improvement yield to avoid having the same problems in subsequent phases present in the project mentioned above.

<sup>11</sup> The project, now considered digital archaeology, uses CAD-based technology for 3D reconstruction. Cologne 3D, therefore, was born as an interdisciplinary project combining digital technical skills with historical-archaeological skills. Firstly, the project reconstructed the “land” where the construction of the city substratum was subsequently built. Data writing follows a standard XML format.

<sup>12</sup> The HallerNet project results from the work inspired and initiated by the doctor-poet von Haller. Haller has been the subject of intense research, especially of a bibliographic nature, since the 19th century. The project could have a follow-up if additional funds are obtained from investors. Thanks to a partnership between the University of Bern and the CCEH, the project was founded, starting from 2014 (reinvestment 2018/2020); considering the high expectations, five applications were opened (including manager, developer, and provider) that could fulfil the needs of enormous work created during the period in question.

<sup>13</sup> (Re)-Collecting Theater History is consequential to work started during the university course of the project coordinator to map the history of the German theatre. Since this is a complex initiative that works on a transversal level and follows a particular trend, it required

site C-Salt<sup>14</sup>[14], Itineranova<sup>15</sup>[15], VedaWeb<sup>16</sup>[16]. The choice of this increase is based on the need not to focus on the observation of projects that could be considered digital archaeology. Hence, the need to expand the case studies to more recent projects that have immediate repercussions or that, in some cases, are still in progress. These examples give us a complete picture of the methods and processes used to create digital cultural artefacts in about a decade. With rare exceptions, all the projects conceived by the CCeH and the IDH<sup>17</sup> have the same development methods. The projects arise mainly following the participation of European calls, including Horizon calls<sup>18</sup>. Hence, the resources obtained<sup>19</sup> are dedicated to creating a particular flow of funding to lay the foundations for the projects.

Sometimes, projects arise at the request of private or public clients, who subsidize the research and development of predetermined solutions to enhance cultural heritage. The funds have a medium-long gestation coinciding with the conclusion of the projects unless a reinvestment of the funds, the projects and digital products made are -apparently- left to fend for themselves.

This attitude generates a significant problem regarding protecting and conserving the digital materials created. The lack of funds, the more consistent coming from the budget of the Horizon projects, is a deterrent to the preservation of digital cultural artefacts. Hence, it is almost impossible to guarantee the protection of digital creations generated without a continuous flow of investments.

Failure to protect and maintain digital creations generates a defect. Therefore, to preserve and safeguard what is not considered in the orbit of the near future, the innovations designed by these projects will no longer be usable. Therefore, with this awareness of a scheduled investment flow, a protection plan should be considered in the project design phase to safe guard the digital cultural material produced.

## 2. DEVELOPMENT AND DESIGN METHODS FOR ACADEMIC RESEARCH

All projects observed, followed and analyzed generally produce numerous data sources, implementing existing processes or, in some cases, trying to make new ones. Undoubtedly, each of these generates digital materials that could fall under the order of digital cultural heritage. However, it is a hypothetical digital cultural heritage (DCH), created and discarded when the concept of conservation, use and consultation decays. Furthermore, the prerequisites previously indicated and cited to define the digital material produced as part of the digital cultural heritage are missing.

For this reason, in an attempt to identify a unique methodology in digitization processes that can be scaled on a larger scale and with a specific non-armored purpose in academic contexts, it would be appropriate to make some considerations.

One of the first assumptions used in German departmental centers involves the identification of a user target to which to dedicate the project created. In most of the projects analyzed, the question is trivialized, considering that the main target is cultural sector practitioners or students. Thus, the end user's needs heavily influence the development choices of the project. However, the definition of a target implies different methodological and theoretical approaches. Therefore, CCeH tries to contact the researchers and others who will use the published resources for each research project. Subsequently, it tries to intercept the actual delivery of editorial output, creating direct contact with the historical source, deep engagement with how they can be represented adequately in the digital medium, collaboration with teams of specialists from related disciplines, and a technological component. In line with the current demands of the digital market, each project is designed and conceived as multi-layered and diverse, integrating many different disciplines and domains.

---

the development of a predominant cultural background on the part of the managers to carry out a decidedly multifaceted work. The value and skill of the coordinator consist in knowing how to manage a wide range of "heritage". The time frame considered is concentrated in seventy years, focusing on a period ranging from 1900 to 1970. The aim is to create a digital archive exploring geo-localized German theatrical history (from Berlin to Cologne). The work is divided between managers and students. In this case, despite the hypothesis of use, the impossibility of exploiting the Agile methodology has materialized. The initial choice of the standard to be applied in the cataloguing fell on the TEI standard, set aside in favor of LIDO (Lightweight Information Describing Objects) -excellent object description tool- considered more responsive to the needs of the work put in place. Among the most pressing problems in the continuation of the project is copyright.

<sup>14</sup> C-Salt was born as a website containing several thematically related projects (including Veda Web). C-Salt is on Beta; it needs to be completed and launched.

<sup>15</sup> Itineranova is a project of the Leuven City Archive. The project aims to digitize and make about 950,000 pages from lay counts of 1362-1795. Digital artefacts are provided with both as scans and in full text with indexing in a portal. In this project, the support of a community created ad hoc to make up for the considerable work behind such a massive digitization process plays a fundamental role.

<sup>16</sup> Veda Web was born as a natural prosecution of Sanskrit online, presenting pros and cons right from the design stage.

<sup>17</sup> Institut für Digital Humanities, Home page, <https://dh.phil-fak.uni-koeln.de>. (last consulted: 04/03/2023).

<sup>18</sup> European Commission, Horizon Europe, [https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe\\_en](https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en). (last consulted: 10/11/2021).

<sup>19</sup> European Commission, Horizon 2020 Framework Programme (H2020), Funding & tender opportunities, <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/programmes/h2020>. (last consulted: 02/03/2023).

Therefore, a successful realization requires an extraordinary and synergetic combination of resources and expertise from established research and knowledge institutions, national libraries and archives, professional publishers, IT specialists and developers from the private sector. After clearly defining the reference target for developing the project, the next phase is concretized. In this fundamental step, the work is organized by specifying the same parcellation following an AGILE method [17]. This choice is governed by the incisiveness and speed with which the various tasks are completed. Like many European innovative projects, multiple operators achieve a single goal (often represented by a WP).

Among the various choices that the CCeH staff typically compile in the design of a project, it is to be hypothesized which standard could be the best choice for the representation of texts in digital form. In creating data, we could cite TEI<sup>20</sup>, which appears to be a common thread in most CCeH projects. Generally, TEI is a standard for data conversion (conversion to write) with pros and cons. Among the most frequent problems highlighted by the administrators (comparison between projects and interviews with the managers of each project) in TEI is the existence of too-detailed specifications in contrast with too-vague licenses. For this reason, CCeH and IDH staff almost always prefer to use TEI vocabulary more; this allows project managers to take advantage of a standardized data language but is highly flexible to the needs of each project.

Often, the case arose in some projects where TEI was not applicable. In these particular cases, where they have materials that had a working history before the project, the centers' developers faced a non-negligible part of the archive material, which has already been catalogued and digitized. So not being able to use TEI had to find a compromise. Therefore, the Meister Eckhart Archive (MEA) catalogues and metadata, still available as Word tables, are now converted to an XML-based template. An archive had to be created from scratch with many problems and compromises. The first compromise was using the text editor and IDE: AtoM (Access to Memory)<sup>21</sup>. Consequently, technically, the signature part is missing from the MEA. Therefore, when it was decided to combine AtoM with signature, it was necessary to modify the AtoM function (different functions create different needs).

Furthermore, one of the significant problems of MEA emerges when the system is updated, the update causes the loss of functions, and therefore, the work is compromised; for example, pieces of writing and meta-adaptation are lost.

In MEA, there was a lack of communication between the developer and the managers, causing a leak. As a consequence, the project failed its goals. Consequently, it is impossible to exploit the project's promises to date. A further problem in this project is the lack of permits due. In addition, copyrights restrict the use of a significant amount of material that cannot be exhibited. Another project where the standard TEI was substituted with a more effective one was (Re)-Collecting Theater History project. This choice is consequential to a specific necessity:

- Digitize and develop a legacy;
- Combine the legacy created in the digital space;
- Explore possible cross-links;
- Examine the various possible links between biography and history.

In this case, considering the need to describe objects, TEI fails in favor of LIDO (Lightweight Information Describing Objects). As a consequence, the process identifies the following needs and priorities:

- digitization of material artefacts such as costumes, masks, paintings and docs;
- identification of German Theater actors;
- identification of main events related to the theatrical representation.

Many of the technologies and methods used are entirely in line with the trends of recent periods. However, as previously pointed out, in some cases, to comply with pre-existing works, a choice had to be made that could combine different needs to achieve a purpose, such as the usability of the DAs created. It can testify to what was asserted with the process used for the Hallernet project. In this case, images with partial metadata writing from the University of Bern (work performed by students, presumably not interpolated images) were used. Creation of a digital archive, presence of texts with file writing in the first instance according to Json principles, then converted into XML language, and subsequent application of the TEI format considered the best choice for merging writing styles.

---

<sup>20</sup> TEI, Text Encoding Initiative, <https://tei-c.org>, (last consulted: 03/03/2023).

<sup>21</sup> Access to Memory (AtoM) Foundation, What is Access to Memory (AtoM)? <https://accessmemoryfoundation.org/access-to-memory/>, (last consulted: 03/03/2023).

### 3. METHODS AND PROCESSES FOR AN ACADEMIC RESEARCH DEVELOPMENT, PROBLEMS AND GOOD PRACTICES

All the projects analyzed have achieved the proposed objectives. Furthermore, each has produced different solutions contextualized to the topic of interest in terms of development. However, every digital result and artefact should be parameterized to the creation period. Reasonably, the goal achieved in the Colonia 3D project is not the same as the more recent projects, both in terms of technologies used for development and in terms of conceiving perspectives that today seem banal concepts such as the protection or safe guarding of the DAs created. However, what is truly interesting in this analysis context are the different method sand approaches used by CCEH and DH to create digital creations. In particular, the German centers have worked on innovation, creating different DAs facing all the trends needed in the digital cultural scenario. Thus, the proposals bounced from the production of 3D reconstruction to the creation of Digital Libraries (DL), apps and web sites. Furthermore, in the creation of the data, the use of the TEI vocabulary has made it possible to compose a flexible and usable markup language according to the requests and needs of the case; this type of “mobility and compositional freedom” can authentically be scalable on multiple contexts and projects and expandable to heterogeneous contexts not linked only to the academic sphere. The projects were born in an academic research and development context; all the creations were preparatory to developing scientific contributions and results according to the classical revision schemes. Generally, one of the most challenging needs that can be countered by analyzing the projects of the departmental centers of the University of Cologne is, as previously stressed, a “failure” to consider a long-term strategy for safe guarding and protecting the digital resources created. However, It is a fact that at least they try to hypothesize a mechanism whereby they can try not to frustrate thoughtful work in which time and money have been invested. The CCEH and IDH to stock and save data use a repository belonging to the center of the University of Cologne. However, this option still creates a disservice due, in the first place, to the possibility of storing the data for a limited time ranging from five to ten years. The limited time of storing means that probably in ten years, the data produced will be lost if they are not preserved, or the preservation of the same is not confirmed. This loss controverts the specifications defined previously in terms of process and memory. Secondly, since it is a service limited to members of the university body or, upon specific request, limited to project managers, not everyone can take advantage of the DAs produced by creating a friction problem and limiting the latter. Furthermore, an ulterior issue arises. Finally, the projects' funds have a significant influence on preservation. Often, if the funds are interrupted or the investment ends, the digital material produced is not followed as it should in the absence of replenishment of money. Thus, running the risk that the hypothetical DCH created will be lost after a short period, more or less a decade (in conjunction with technological obsolescence and system outages we are increasingly going against). However, this issue is not a problem that belongs to the CCEH only but is expansible to a macrocosm. Moreover, this criticality emerged is scalable across multiple contexts and areas, both academic and industrial, becoming, as expected, a topic of discussion. This issue needs a double solution as quickly as possible, with results applicable in the medium and long term.

#### REFERENCES

- [1] Ciotti, Fabio. 2018. «Dall'Informatica umanistica alle Digital Humanities. Per una storia concettuale delle DH in Italia.» Digital Humanities 2018. Puentes-Bridges. Book of Abstracts. Libro de resúmenes. 174-177.
- [2] McGann, Jerome J., Memory Now, «4Humanities», 19 August 2012, <http://4humanities.org/2012/08/jerome-j-mcgann-memory-now-2/>, (last consulted: 04/05/2023).
- [3] Garrett, John R. 1995. «Task Force on Archiving of Digital Information.» D-Lib Magazine.
- [4]Singh, Anil. 2012. "Digital preservation of cultural heritage resources and manuscripts: An Indian government initiative." International Federation of Library Associations and Institutions 38 (4): 289-296.
- [5] Tomasi, Francesca. 2017. "La preservazione del contenuto degli oggetti culturali: formalizzare la provenance." BIBLIOTHECAE.IT 25.
- [6] Däßler, Rolf, and Ulf Preuß. 2020. "Digital Preservation of Cultural Heritage for Small Institutions." In Digital Cultural Heritage, by Horst Kremers, 109. Switzerland: Springer.
- [7] DiXiT,Homepage, DigitalScholarlyEditionsInitial TrainingNetwork,<http://dixit.uni-koeln.de>,(last consulted:04/03/2023).
- [8] Philosophische Fakultät Thomas-Institut, Meister-Eckhart-Archiv, Universität zu Köln <https://thomasinstitut.uni-koeln.de/forschung/arbeitsstellen-und-langzeitprojekte/meister-eckhart-archiv> (last consulted: 03/03/2022).
- [9] Funderburk, Jim, Patel, Dhaval, Cologne Digital Sanskrit Dictionaries, Universität zu Köln <https://www.sanskrit-lexicon.uni-koeln.de> (last consulted: 03/03/2023).
- [10] Lateinische Paläographie. Der Antike und des Mittelalters, Paläographie Online, <http://www.palaographie-online.de> (last consulted: 03/03/2023).

- [11] Colonia3D, Visualisierung des Römischen Köln, <http://colonia3d.de/colonia3d-home/> (last consulted: 03/03/2023).
- [12] HallerNet, Home page, <https://hallernet.org/data/person/01200/core> (last consulted:15/11/2021).
- [13] RE-CollectingTheaterHistory, (Beta0.1.4), <https://recollecting.tws.uni-koeln.de/index.html> (last consulted: 15/11/2021).
- [14] C-Salt, Cologne South Asian Languages and. Texts, <http://c-salt.uni-koeln.de> (last consulted: 04/03/2023).
- [15] Itineranova, Homepage, <https://www.itineranova.be/in/home> (last consulted: 04/03/2023).
- [16] VedaWeb, Online Research Platform for Old Indic Texts, <https://vedaweb.uni-koeln.de> (last consulted: 04/11/2021).
- [17] Huff, Sid, Diane E. Strode, and Alexei, Tretiakov 2009."The Impact of Organizational Culture on Agile Method Use".42nd Hawaii International Conference on System Sciences. IEEE. 1-9.



# Preservare il *Tesoro*.

## Per un nuovo sistema di interrogazione dei corpora dell'Opera del Vocabolario Italiano

Emiliano Degl'Innocenti<sup>1</sup>, Leonardo Canova<sup>2</sup>, Francesco Coradeschi<sup>3</sup>, Federica Spinelli<sup>4</sup>

<sup>1</sup> DARIAH-IT National Coordinator, Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia - emiliano.deglinnocenti@cnr.it

<sup>2</sup> Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia – canova@ovi.cnr.it

<sup>3</sup> Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia – coradeschi@ovi.cnr.it

<sup>4</sup> Istituto Opera del Vocabolario Italiano (CNR-OVI), Italia – spinelli@ovi.cnr.it

### ABSTRACT

Questo articolo intende presentare il progetto per un nuovo sistema di interrogazione dei corpora testuali dell'Opera del Vocabolario Italiano, nato dalla collaborazione tra l'Istituto e il nodo italiano dell'infrastruttura di ricerca europea DARIAH e volto a fornire, con un approccio Open Science, una piattaforma d'accesso *user friendly*, accessibile e sostenibile ai corpora, che rappresentano ad oggi la base testuale sulla quale si definisce il concetto di "italiano antico". Ad una necessaria introduzione relativa al software GATTO, da 25 anni alla base di tutte le operazioni lessicografiche svolte all'interno dell'Istituto, segue la definizione degli scopi e dei requisiti richiesti, la concettualizzazione dello strumento e del suo rapporto con GATTO e la presentazione delle prime fasi di sviluppo, relative alle funzioni di ricerca per forme e per lemmi, e dei primi risultati. Si offre infine un essenziale cronoprogramma e una panoramica sull'inserimento del progetto nel framework europeo dell'Open Science attraverso i progetti PON e PNRR di cui DARIAH è promotrice.

### PAROLE CHIAVE

Lessicografia; Digital Libraries; italiano antico; database; sostenibilità

## 1. L'OPERA DEL VOCABOLARIO ITALIANO E GATTO

Il tedesco *Wortschatz* ("vocabolario") è una forma composta dalle parole *Wort* ("parola") e *Schatz* ("tesoro"), che può tradursi alla lettera come "tesoro di parole", "patrimonio linguistico". L'Opera del Vocabolario Italiano (da qui in avanti semplicemente OVI)<sup>1</sup> nasce nel 1985 come Centro di studi, per costituirsi come Istituto del Consiglio Nazionale delle Ricerche (CNR) nel 2001. L'OVI è ad oggi il principale agente della preservazione, interpretazione e disseminazione del "tesoro" dell'"italiano antico", concetto sfuggente che – grazie alle sue risorse – lo stesso istituto ha contribuito a definire in maniera sostanziale.<sup>2</sup> L'OVI ospita infatti la redazione del *Tesoro della Lingua Italiana delle Origini (TLIO)*, pubblicato online a partire dal 1997 e comprendente – ad oggi (aprile 2023) – 44003 voci, che coprono «tutte le varietà italo-romanze di cui sono rimaste testimonianze medievali» [4.756]. Alla base di questa monumentale opera sta il *Corpus TLIO*, una banca dati digitale creata appositamente contenente tutti i testi italiani composti entro il 1375 – anno della morte di Giovanni Boccaccio – disponibili in edizioni considerate affidabili. Queste, nel loro insieme, rappresentano oggi un canone che, nella percezione degli studiosi di settore, è quanto di più vicino al concetto di "italiano antico" di cui sopra.<sup>3</sup>

Da circa venticinque anni, questo straordinario tesoro è gestito dal software GATTO (*Gestione degli Archivi Testuali del Tesoro delle Origini*), sviluppato da Domenico Iorio-Fili con la collaborazione di altri ricercatori dell'OVI e oggi gestito da Andrea Boccellari. Si tratta di uno strumento che mette a disposizione degli utenti diversi tool – essenziali per soddisfare le necessità dell'Istituto – che vanno dalla costruzione e gestione del corpus, alle operazioni di POS-tagging e lemmatizzazione dei testi, fino all'interrogazione del corpus lemmatizzato [5.376], che viene attualmente impiegato nella sua release 3.3, scritta in linguaggio Visual Basic 6.0 e destinata ad operare in locale su ambienti Windows<sup>4</sup>. I corpora allestiti con GATTO sono consultabili liberamente online attraverso la piattaforma [GattoWeb](#), che costituisce la versione web (limitata alla sola funzione di interrogazione dei testi) del software GATTO 3<sup>5</sup> e rappresenta l'endpoint attraverso il

<sup>1</sup> Cf. lo [statuto](#) dell'OVI.

<sup>2</sup> Cf. [4.761].

<sup>3</sup> Il *Corpus TLIO* contiene (al 2 ottobre 2022) 3173 testi, per un totale di 23685634 occorrenze, ed è il corpus su cui il TLIO viene redatto di prima mano. Altri corpora, legati a specifici progetti, sono consultabili sul [sito dell'OVI](#).

<sup>4</sup> Il linguaggio di programmazione Visual Basic 6.0 è progettato per operare su sistemi operativi a 32 bit. Dal momento che la maggior parte delle macchine oggi opera a 64 bit, GATTO 3.3 può essere eseguito soltanto su macchina virtuale o su hardware datato.

<sup>5</sup> Su GattoWeb e sul suo rapporto con GATTO vedi anche la relativa [guida](#).

quale è possibile accedere alle risorse testuali con funzioni e viste specificamente pensate per il dominio della lessicografia: ricerche per forma, per lemmi, per co-occorrenze- con possibilità di impiegare metacaratteri (*wildcards*, caratteri jolly) e di definire sottocorpora. Si tratta, a ben vedere, della più estesa e autorevole *digital library* – pensata specificamente per la lessicografia dell’italiano antico – che, se di certo non rispecchia il sogno espresso da H. G. Wells nel 1938 di un «complete planetary memory for all mankind» [7.XXVIII], può forse rientrare nella definizione – meno altisonante ma certo soddisfacente – di «an (almost) complete ancient italian language memory for all mankind». Attorno a tale riflessione teorica, tuttavia, sorgono inevitabilmente domande più pratiche: quali garanzie di sostenibilità abbiamo nel futuro prossimo e meno prossimo [11]? Quali fondi e quali infrastrutture saranno necessarie [10]? Com’è possibile favorire la transizione di questo materiale verso l’Open Science [8; 9]? Può una tale transizione incrementare il (già enorme) valore di questo patrimonio per la società [1; 7]?

## 2. PER UN NUOVO SISTEMA DI INTERROGAZIONE DEI CORPORA OVI

Da tali riflessioni e preoccupazioni sorte all’interno dell’Istituto, che coordina per il CNR le attività del [nodo nazionale italiano di DARIAH-ERIC](#), e in seno all’esperienza maturata nell’ambito del [Progetto RESTORE](#) [2; 3] è nata la proposta di realizzare un nuovo sistema di interrogazione per i corpora dell’OVI. Dal confronto fra i ricercatori dell’OVI ed esperti di settori diversi – lessicografi, linguisti, filologi e informatici per le scienze umane sono emerse le seguenti priorità, che ne hanno guidato lo sviluppo:

1. preservare i dati dei corpora OVI dal rischio di obsolescenza dei software attualmente utilizzati per l’interrogazione;
2. progettare un sistema di interrogazione *user friendly* fruibile anche da parte di utenti privi di competenze specifiche nel campo della lessicografia;
3. guidare la progettazione e lo sviluppo di funzioni innovative e garantire la piena interoperabilità con altri sistemi interni ed esterni all’OVI.

Conseguentemente, sono stati definiti i requisiti tecnici della nuova piattaforma:

1. consentire la ricerca per forme e per lemmi anche grazie all’impiego di caratteri *wildcard*;
2. consentire il download degli esempi individuati mediante le ricerche in diversi formati (e.g. PDF, DOC, DOCX, RTF, XML-TEI), con possibilità di personalizzazione dell’ampiezza dei contesti, e facilitarne la possibilità di citazione bibliografica;
3. consentire la ricerca per co-occorrenze e auspicabilmente incrementarne l’efficienza rispetto al software GATTO;
4. consentire la definizione di un sottoinsieme di testi in cui effettuare le ricerche, definibile in base a parametri inseriti nella scheda bibliografica del testo stesso (l’area linguistica, il nome dell’autore, la data del testo, il genere d’appartenenza, ecc.).

Il nuovo software nasce inizialmente con l’obiettivo di replicare il funzionamento della piattaforma GATTO, utilizzando tecnologie e soluzioni *up-to-date*. Tuttavia, il suo scopo principale è quello di fornire agli utenti la possibilità di interrogare una varietà di corpora lessicografici, attraverso un’interfaccia intuitiva e *user-friendly*, fruibile anche da parte di utenti con conoscenze di dominio non approfondite, senza compromettere la correttezza e affidabilità dei risultati di ricerca.

### 2.1. Concettualizzazione ed esigenze specifiche

Come si è visto sopra, ad oggi l’Istituto gestisce i suoi corpora tramite la release 3.3 di GATTO, che soddisfa pienamente le esigenze scientifiche dell’Opera del Vocabolario, e ha un suo endpoint nella piattaforma GattoWeb. Di GATTO, tuttavia, esiste una versione successiva (4.0) rilasciata nel 2010 dallo stesso Iorio-Fili, che andava ad integrare nuove funzioni e a risolvere alcune criticità: veniva infatti introdotta una funzione sperimentale di lemmatizzazione semi-automatica [6], incrementata l’efficienza e l’affidabilità della ricerca per co-occorrenze e – fatto di interesse fondamentale ai fini del nostro progetto – compiuto uno sforzo verso l’impiego di linguaggi di codifica standard e, dunque, verso un principio di interoperabilità. In particolare, GATTO 4.0 fa ricorso a [SQLite](#) per la gestione dei database (precedentemente Microsoft Access), implementa il set di caratteri Unicode appartenenti al Basic Multilingual Plane (affiancati ai caratteri ANSI, già introdotti nella release precedente) e, soprattutto, introduce la possibilità di inserire nei corpora testi marcati in linguaggio XML, riuscendo a gestire un subset di tag XML-TEI, unitamente al sistema di marcatura proprietario e - dunque - non

standard. Negli anni scorsi, tuttavia, dopo una lunga fase di test questa versione del software è stata quasi completamente abbandonata dall'Istituto, dal momento che GattoWeb non gestisce i corpora allestiti con GATTO 4.0 e che un endpoint specifico per questa release non è stato portato a compimento.

A partire da questo elemento mancante, dunque, la concettualizzazione del nuovo sistema cerca di tenere conto da un lato delle esigenze dell'Istituto di continuare a lavorare su un software per il quale ci sono competenze sviluppate e programmi di addestramento<sup>6</sup>, dall'altra di aprirsi al mondo dell'Open Science nell'ottica di offrire – con strumenti all'avanguardia – i risultati del lavoro di ricerca ad un pubblico più ampio rispetto alla cerchia ristretta di lessicografi, linguisti e filologi. Date queste premesse, dunque, si è deciso di contravvenire almeno in parte al principio enunciato da Iorio-Fili in favore di uno strumento unico che consentisse di svolgere tutte le operazioni, dalla redazione alla visualizzazione dei corpora «con il duplice scopo di utilizzare al meglio le ovvie sovrapposizioni funzionali esistenti [...] e di migliorare l'affidabilità complessiva»[5.367], in virtù di un approccio modulare maggiormente conforme ai criteri di sviluppo contemporanei, e conseguentemente di sviluppare uno strumento *stand-alone* che, a partire da corpora allestiti in ambiente GATTO, offra una visualizzazione dei dati dinamica, *user friendly* e adatta a un pubblico più ampio.

## 2.2. Workflow e prime implementazioni

Le operazioni di generazione e editing (lemmatizzazione, POS-tagging, annotazione ecc.), ossia tutto ciò che avviene internamente all'Istituto, rimarranno dunque dominio esclusivo di GATTO 3.3, così come la piattaforma GattoWeb, per i motivi di sopra enunciati, resterà la piattaforma preferenziale per la realizzazione delle voci del TLIO. Del corpus allestito in GATTO 3.3 sarà poi creata una copia, convertita in formato GATTO 4 tramite lo strumento di conversione presente nel software stesso, ed è su questa copia che opererà il nuovo sistema di interrogazione, andando dunque non a sostituire, ma ad affiancare GATTO. Quest'ultimo, diversamente dalla maggior parte dei software che permette di effettuare ricerche su corpora testuali, non opera scorrendo i testi codificati e confrontando i parametri immessi con quelli incontrati, bensì quando «isola e riconosce le parole, le interpunzioni, le frasi e altri elementi che lo costituiscono e provvede a registrare tali elementi in apposite tabelle all'interno di un database: quando un utente interrogherà il corpus, la richiesta verrà analizzata per riconoscerne la tipologia, individuare gli oggetti cercati e le caratteristiche dei contesti voluti (dimensioni o altro), dopo di che la ricerca verrà effettuata all'interno delle tabelle del database. Queste contengono degli indici che permettono di localizzare rapidamente le informazioni cercate; in modo analogo si ottengono le informazioni di natura bibliografica o testuale relative ai contesti (titolo e autore dell'opera; pagina, riga e altre coordinate all'interno del testo)» [5.373]. Alla base della ricerca, dunque, non vi sono i testi ma le tabelle del database generato da GATTO – in formato SQLite nella versione 4.0 – che contengono gli indici di rimando alle posizioni specifiche nei testi.

### 2.2.1. Motore di ricerca

Il nuovo sistema di interrogazione, dunque, si avvale di un motore in linguaggio di programmazione [Python](#) che crea dinamicamente le *queries* SQLite in base all'input dell'utente. Il sistema include un subset delle funzioni di ricerca presenti in GATTO – ritenuto adeguato per rappresentare le esigenze dell'utente di riferimento e soddisfare i requisiti specificati in precedenza – oltre che la possibilità di implementare ulteriori funzioni e viste sui dati. Dal momento che GATTO ammette caratteri di interrogazione non-standard (“\*” per zero o più caratteri, “?” per un solo carattere, “<” per caratteri alternativi separati da virgole, ad es. “<a, e>”) è sorta la necessità di sviluppare una funzione interprete, che permette di tradurre tali caratteri nei corrispondenti caratteri ammessi dalla sintassi SQLite (rispettivamente “%”, “\_” e una lista di opzioni alternative per “<”). La scelta di isolare l'interprete in una funzione specifica permette quindi di mantenere standard il linguaggio di interrogazione del database, concentrando al suo interno l'elaborazione dei parametri non standard, a tutto vantaggio degli utenti abituati a lavorare con GATTO. L'interprete consentirà inoltre di gestire la cifratura del database e dei testi, qualora l'Istituto decida di mantenerla all'interno del nuovo sistema.

Conseguentemente, si è passati a riprodurre le funzioni di ricerca per forme e per lemmi in linguaggio Python, che operano rispettivamente sulle tabelle FORM e LEM del database generato da GATTO. Queste ammettono diverse opzioni, tra le quali:

- ricerca per iniziale raddoppiata, che inserisce tra i valori della clausola WHERE della query il valore immesso con il primo carattere raddoppiato (utile per rilevare i casi di raddoppiamento fonosintattico);

---

<sup>6</sup> L'Opera del Vocabolario Italiano organizza annualmente corsi di formazione per la lessicografia dell'italiano antico, nei quali i partecipanti vengono istruiti sull'utilizzo di GATTO per la redazione delle voci del TLIO.

- ricerca espansa, che include nella ricerca tutti i caratteri speciali e accentati, operando a partire dalla colonna SPEC della tabella (contenente la forma con i caratteri speciali) per recuperare la forma normalizzata nella colonna NORM, grazie alla quale recuperare poi tutte le forme ad essa associate;
- vista lemmi/forme, che mostra i lemmi associati alle forme nel caso della ricerca per forme o le forme associate ai lemmi nel caso della ricerca per lemmi, incrociando le due tabelle tramite l'operatore INNER JOIN.

Tali funzioni di base sono poi combinate in una funzione di ricerca per co-occorrenze, che permette di inserire più query di ricerca – per ognuna delle quali è possibile specificare i parametri (forma/lemma, iniziale raddoppiata, ricerca espansa) – e di specificare la distanza tra le parole cercate all'interno del testo, la volontà di cercare le parole all'interno di uno stesso periodo o di cercarle esattamente nell'ordine in cui sono state inserite. A partire da una normale ricerca fatta sulla prima delle richieste inserite, la funzione lancia, all'interno della prima lista dei contesti, la seconda delle richieste (secondo i parametri inseriti) e sottrae dalla lista dei risultati tutti i contesti che non soddisfano la seconda richiesta, e procede in questo modo per ognuna delle parole immesse, fino a raggiungere la lista finale dei contesti che contengono le co-occorrenze richieste. Le funzioni di ricerca restituiscono i risultati, ordinati alfabeticamente secondo l'indice IDFOR/IDLEM presente nella tabella, all'interno di un *dataframe* [Pandas](#), facilmente manipolabile ed esportabile in molteplici formati, quali JSON e CSV. A partire da questo *dataframe*, che contiene gli elementi localizzati nel database e le informazioni ad essi collegate, due successive funzioni elaborano i dati contenuti in altre tabelle del database, permettendo dapprima di individuare i testi in cui tali elementi sono contenuti e in seguito il puntatore al primo carattere della forma nei testi. Il risultato dell'elaborazione si configura come un nuovo *dataframe* contenente una lista di *snippet* associati alla forma o al lemma immessi in fase di ricerca in set di caratteri Unicode, che un'altra funzione incrocia con il database – indipendente da quello che contiene il corpus – in cui sono inserite le informazioni bibliografiche. Per mezzo di un ulteriore set di funzioni, poi, è possibile estrarre i singoli contesti, associare le note (prevalentemente apparati critici) e i brani associati (prevalentemente testi latini dei volgarizzamenti) e personalizzare l'ampiezza del contesto per numero di parole, di periodi e di brani associati secondo le informazioni reperibili nel database.

I test per la valutazione dell'efficacia del nuovo sistema di interrogazione sono stati svolti su un sottocorpus del *Corpus TLIO* di dimensioni ridotte, composto da testi considerati rappresentativi di alcuni fenomeni caratterizzanti i testi in italiano antico e, al netto di minime difformità dovute alla configurazione del corpus di prova, hanno restituito risultati identici a quelli forniti da gatto, a testimonianza del buon esito del lavoro di porting.<sup>7</sup>

### 2.2.2. Struttura della piattaforma

La piattaforma usa un'architettura a servizi: è composta cioè di più componenti discreti sviluppati separatamente, che comunicano tra di loro tramite API REST che accettano e restituiscono specifici *data transfer objects*, strutturati secondo le necessità dei singoli servizi.

Il “cuore” della piattaforma è costituito da una serie di servizi di *back-end*, che interpretano le richieste (oggetti in formato JSON) che arrivano dalle interfacce utente, restituendo strutture di dati (sempre in JSON) che sono quindi usate per popolare le interfacce di visualizzazione (e che possono, se necessario, anche essere visualizzate direttamente in forma testuale). Queste parti contengono tutta la logica della piattaforma (motori di ricerca, interpreti e trasformatori di dati) e sono scritte – come si è visto – in linguaggio Python. Per creare gli *endpoint* necessari ad esporre le API come servizi web si utilizza il micro-framework Flask<sup>8</sup> ed il server nginx<sup>9</sup>; per ragioni di sicurezza e ottimizzazione delle prestazioni, quest'ultimo è situato dietro un server Apache<sup>10</sup> che agisce come *reverse proxy* per filtrare e gestire le richieste che provengono dall'esterno. È opportuno sottolineare che l'accoppiamento tra questi componenti tecnologici è debole (i.e.: *loose coupling*): le funzioni logiche non sono dipendenti dalle caratteristiche di Flask (e possono essere usate anche direttamente come *script* Python senza doverli integrare in un servizio; questa caratteristica può essere utile in diversi contesti ed è usata in fase di sviluppo e per lo *unit testing*). D'altra parte gli *endpoint* generati da Flask possono essere integrati, senza sforzo aggiuntivo, con server diversi da nginx ed Apache. Questa indipendenza tecnologica dei componenti è studiata per favorire la futura manutenibilità e scalabilità della piattaforma.

La piattaforma è configurata per poter usare multipli database (DB), ognuno dei quali gestito da un servizio dedicato e tecnologicamente indipendente dagli altri. In particolare, il DB che contiene le informazioni su forme, lemmi e contesti è mutuato da GATTO (in particolare, nella versione 4.0), ed ha una struttura ibrida, composta da una serie di file SQLITE

<sup>7</sup> L'accuratezza dell'output è stata calcolata eseguendo ricerche con i due sistemi, inserendo in input gli stessi dati.

<sup>8</sup> Flask è un micro-framework Web scritto in Python. <https://flask.palletsprojects.com/en/2.3.x/>.

<sup>9</sup> Nginx, è un web server open source, utilizzato anche come proxy inverso. <https://www.nginx.com/>.

<sup>10</sup> Apache HTTP Server è un web server open-source. <https://httpd.apache.org/>.

(usati per lemmi, forme, occorrenze, bibliografia...) e da una collezione di documenti di testo (utilizzati per i contesti e per la loro formattazione); l'accesso ai dati conservati nel DB avviene mediante un servizio realizzato con le stesse tecnologie del *back-end*: Python, Flask, nginx.

Come prossimo sviluppo, è prevista l'integrazione nella piattaforma di un altro DB (già sviluppato), contenente *dati semantici*, al fine di arricchire il contesto in cui opera il TLIO. Tale aggiunta consentirà di integrare metadati e dati aggiuntivi relativi agli autori, alle edizioni dei testi, al contesto storico, geografico e altre informazioni pertinenti. L'integrazione consentirà inoltre la connessione con altre applicazioni, ad esempio un software per la visualizzazione di Digital Scholarly Editions, come EVT<sup>11</sup>. Dal punto di vista tecnologico, questo servizio è implementato tramite il *triple store* Virtuoso<sup>12</sup>, e la comunicazione con gli altri servizi gestita tramite l'*endpoint* SPARQL messo a disposizione da Virtuoso.

La struttura a servizi che caratterizza la nuova piattaforma, garantendo la separazione dei dati dalle interfacce, rende possibile l'implementazione di front-end personalizzati secondo le esigenze di diversi casi d'uso e/o gruppi di utenti che, pur consumando i dati provenienti dai servizi di back-end, risultano del tutto indipendenti dal punto di vista tecnologico e gestionale. La prima interfaccia di visualizzazione, tutt'ora in fase di completamento, è stata sviluppata per la gestione della fase di test dell'applicazione da parte dei ricercatori dell'OVI e sarà ulteriormente raffinata, attraverso un processo di *cooperative design*, prendendo in considerazione i requisiti di accessibilità e *user friendliness* di cui sopra. Si tratta di un'applicazione web scritta in [HTML5](#) e Javascript (avvalendosi della libreria [jQuery](#) per la gestione delle richieste agli altri servizi), servita tramite Apache, il cui design, è volutamente minimale, così come limitato al minimo è l'uso di pacchetti e librerie esterne. In particolare, gli utenti della piattaforma GattoWeb hanno spesso lamentato la necessità di dover attraversare diversi passaggi (ricerca, elementi localizzati, accumulatore, contesti multipli, contesti singoli) per accedere ai risultati delle loro ricerche; pertanto, la nuova interfaccia andrà a caratterizzarsi quanto più possibile come una *single-page application*, al fine di ridurre al minimo o eliminare del tutto i passaggi intermedi per fornire un'esperienza utente più rapida e fluida.

### 2.3. Futuri sviluppi e risultati attesi

L'interfaccia di visualizzazione attualmente in corso di sviluppo, da intendersi come prima applicazione del software ad un concreto caso d'uso, andrebbe dunque a configurarsi come endpoint per GATTO 4.0 (come GattoWeb per GATTO 3.3) offrendo una serie di vantaggi sia per i ricercatori dell'Istituto, sia per gli utenti esterni. Quanto ai primi, il principale requisito espresso è stato relativo allo sviluppo di un'interfaccia di interrogazione online in grado di visualizzare i corpora prodotti con GATTO 4.0 in maniera snella ed efficace, a supporto della transizione dell'Istituto verso il nuovo software: più performante, tecnologicamente aggiornato, più aperto agli standard internazionali e maggiormente aderente all'approccio FAIR ai dati della ricerca. Rispetto ai secondi, poi, la concettualizzazione dell'interfaccia in una prospettiva aperta al pubblico con esigenze non (soltanto) lessicografiche, garantirà plausibilmente una diffusione più capillare dei contenuti dei corpora OVI e di conseguenza un maggiore impatto nel mondo scientifico e nella società, in termini di definizione del canone e del concetto di "italiano antico". Inoltre, lo sviluppo di un sistema *stand-alone* ma appoggiato a GATTO 4.0 permetterà, da un lato, di accedere ai dati in formato standard XML-TEI e – dall'altro, in un'ottica di riutilizzo – di poterlo eventualmente adattare all'interrogazione di altri corpora testuali.

Il team di sviluppo è attualmente al lavoro sulla parte sistemistica del progetto, sul raffinamento dell'interfaccia e sulle ultime funzioni da implementare, quali la definizione di sottocorpora sulla base di dati bibliografici (autori, date, aree linguistiche ecc.) e il download dei contesti nei formati richiesti. La fase di riscrittura e porting di GATTO, attualmente in fase di completamento, dovrebbe concludersi entro l'autunno del 2023 per lasciare spazio ad attività di personalizzazione dell'interfaccia e di sviluppo dell'applicazione secondo i paradigmi di Open Science e FAIRness dei dati.

Quanto al codice, il software, caratterizzato da una struttura modulare, è stato progettato come applicazione complementare a GATTO 4.0, ma è di fatto un componente *stand-alone*, riusabile in contesti differenti e adattabile ad altri corpora e database. In questo modo, il sistema può essere considerato come una soluzione flessibile e scalabile per l'analisi e l'interrogazione di dati testuali (e in special modo lessicografici) in grado di soddisfare le esigenze di un ampio ventaglio di utenti. In vista della pubblicazione del sistema attraverso l'infrastruttura DARIAH-IT, sarà necessario definire la modalità di rilascio del software, privilegiando modelli di licenza aperta, al fine di aumentare il potenziale di riutilizzo e sviluppo da parte della comunità scientifica di riferimento.

Quanto ai dati, il progetto si colloca all'interno delle iniziative promosse da OVI verso la standardizzazione e la pubblicazione delle sue risorse, in cui rientra anche lo Spoke 3 *Digital libraries, archives and philology* del progetto PNRR

<sup>11</sup> EVT – Edition Visualization Technology. <http://evt.labcd.unipi.it/>

<sup>12</sup> Virtuoso Universal Server. <https://virtuoso.openlinksw.com/>



*CHANGES (Cultural Heritage Active Innovation for Nex-Gen Sustainable Society)*, tra i cui scopi vi è la realizzazione di una *digital library* liberamente consultabile dei corpora testuali gestiti dall'Istituto. In questo contesto, la piattaforma che è oggetto di questo contributo dovrà auspicabilmente muoversi verso i principi FAIR garantendo:

- *findability*, dal momento che ad ogni contesto sarà associato un identificatore permanente che ne faciliti il reperimento e la citazione;
- accessibilità, dal momento che l'API garantirà un accesso intuitivo ai dati e che le modalità di accesso (comprese eventuali modalità di autorizzazione e autenticazione) saranno chiaramente esplicitate nella documentazione;
- interoperabilità, data la sua architettura modulare, potrà offrire il supporto alla combinazione e utilizzo di diversi formati di dati e strumenti. Inoltre, l'utilizzo di formati standard (come XML-TEI) è favorito e facilitato dall'integrazione all'interno del software di uno strumento configurabile, appositamente sviluppato per la trasformazione dei dati e metadati da un formato proprietario al formato standard;
- riusabilità, perché sarà possibile reimpiegare i dati estratti dall'API in altri strumenti software per esigenze di dominio diverse (NLP, lemmatizzatori, NER ecc.).

### 3. SOSTENIBILITÀ E FRAMEWORK EUROPEO

La realizzazione di questo progetto si avvale del supporto dell'infrastruttura di ricerca europea DARIAH-ERIC, che fornisce servizi informatici per le scienze umane. Al progetto, DARIAH offrirà l'hosting all'interno di uno dei datacenter ad alte prestazioni nati in seno al progetto PON *DARIAH-it (Developing national and Regional Infrastructural nodes of dAriaH in Italy (DARIAH.IT) - PIR01\_00022)*, garantendo supporto finanziario, tecnico e scientifico nel lungo periodo. I risultati del progetto entreranno inoltre a far parte del progetto PNRR *H2IOSC (Humanities and Heritage Open Science Cloud)*, all'interno del quale, nell'ambito del WP 7 dedicato a *Community pilots: innovative cross-domain services and environments*, DARIAH propone il *pilot Digital Philology HUB*, dedicato all'interoperabilità tra strumenti dedicati all'analisi del linguaggio, dei testi, dei manoscritti e delle edizioni a stampa, auspicabilmente influenzando sul rapporto che tradizionalmente pone la lessicografia storica in subordinazione rispetto alla filologia [4] e proiettando gli straordinari risultati della ricerca dell'Istituto nel framework internazionale dell'Open Science.

### 4. RINGRAZIAMENTI

Si ringraziano Maurizio Sanesi e Alessia Spadi, che appartengono al team di sviluppo di questo progetto ma che non compaiono tra gli autori di questo contributo; Andrea Boccellari per il supporto nell'interpretazione e nella traduzione del software GATTO; Diego Dotto, Cristiano Lorenzi Biondi e Zeno Verlatto, per la consulenza lessicografica e filologica e Paolo Squillaciotti per l'iniziativa di questo progetto.

### BIBLIOGRAFIA

- [1] Arms, William Y. 2000. *Digital Libraries*. Cambridge, Massachusetts: The MIT Press.
- [2] Degl'Innocenti, Emiliano, Carmen Di Meo, Alessia Spadi, Federica Spinelli, Francesco Coradeschi, e Maurizio Sanesi. 2022. «RESTORE: Opening Data in Digital Humanities and Cultural Heritage». In *Proceedings of the 18th Italian Research Conference on Digital Libraries*. Padua, Italy, February 24-25, 2022 (hybrid event), a cura di Giorgio Maria Di Nunzio, Beatrice Portelli, Domenico Redavid, e Gianmaria Silvello. <http://ceur-ws.org/Vol-3160/>
- [3] Degl'Innocenti, Emiliano, Alessia Spadi, e Federica Spinelli. 2021. «RESTORE. smaRt accESs TO digital heRitage and mEmory». In *Proceedings of the 17th Italian Research Conference on Digital Libraries*. Padua, Italy (virtual event due to the Covid-19 pandemic) February 18-19, 2021, a cura di Dennis Dosso, Stefano Ferilli, Paolo Manghi, Antonella Poggi, Giuseppe Serra, e Gianmaria Silvello, 178–85. <https://ceur-ws.org/Vol-2816/>
- [4] Guadagnini, Elisa. 2016. «Lessicografia, filologia e “corpora” digitali: qualche considerazione dalla parte dell'OVI». *Zeitschrift für romanische Philologie* 132 (3): 755–92.
- [5] Iorio-Fili, Domenico. 2007. «Breve storia, stato attuale e prospettive del software GATTO». *Bollettino dell'Opera del Vocabolario Italiano XII*: 365–86.
- [6] ———. 2010. «Un nuovo strumento di lemmatizzazione automatica per corpora testuali di ridotte dimensioni. Applicazione all'italiano antico». *Bollettino dell'Opera del Vocabolario Italiano* 15: 367–91.
- [7] Lesk, Michael. 2005. *Understanding Digital Libraries*. San Francisco: Morgan Kaufman.

- [8] Macgregor, George, a c. di. 2008. «Digital Libraries and the Semantic Web: Context, Applications and Research». *Library Review* 57 (3).
- [9] Manghi, Paolo, Leonardo Candela, e Gianmaria Silvello, a c. di. 2019. *Digital Libraries: Supporting Open Science: 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 – February 1, 2019, Proceedings*. Vol. 988. *Communications in Computer and Information Science*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-030-11226-4>
- [10] Silvello, Gianmaria, Oscar Corcho, Paolo Manghi, Giorgio Maria Di Nunzio, Koraljka Golub, Nicola Ferro, e Antonella Poggi, a c. di. 2022. *Linking Theory and Practice of Digital Libraries: 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20–23, 2022, Proceedings*. Vol. 13541. *Lecture Notes in Computer Science*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-031-16802-4>
- [11] Skinner, Katherine, e Martin Halbert. 2008. *Strategies For Sustaining Digital Libraries-Emory University Digital Library Publications*. Atlanta: Emory University.



# La galassia MQDQ tra innovazione e tradizione

Angelo Mario Del Grosso<sup>1</sup>, Martina Venuti<sup>2</sup>, Federico Boschetti<sup>3</sup>, Luigi Tassarolo<sup>4</sup>

<sup>1</sup> Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILC), Italia – angelo.delgrosso@ilc.cnr.it

<sup>2</sup> Università Ca' Foscari Venezia, Italia – martina.venuti@unive.it

<sup>3</sup> Istituto di Linguistica Computazionale 'A. Zampolli' (CNR-ILC), Italia – federico.boschetti@ilc.cnr.it

<sup>4</sup> Università Ca' Foscari Venezia, Italia – luigi.tassarolo@gmail.com

## ABSTRACT

MQDQ-galaxy è il nome assegnato a un progetto di ampio respiro, sviluppato dal 2019 e tuttora in corso, volto a rinnovare gli archivi di MQDQ (www.mqdq.it) sia tramite l'incremento di testi e apparati critici, sia tramite una nuova piattaforma di editing collaborativo (Wiki-MQDQ), sia tramite la realizzazione di un modello sostenibile e moderno di gestione e sviluppo degli archivi elettronici di testi. Nel presente contributo si delineano le principali azioni intraprese durante i lavori di MQDQ-galaxy, sviluppati presso il VeDPH (DSU, Ca' Foscari Venezia) e CNR-ILC Pisa.

## PAROLE CHIAVE

Musisque Deoque, Archivi Digitali Latini, filologia digitale, API, intertestualità

## 1. INTRODUZIONE: IL PROGETTO MQDQ-GALAXY E I SUOI CORPORA

Il progetto MQDQ-Galaxy nasce nel corso del 2019 con lo scopo di creare un modello sostenibile e aggiornato per la produzione, la gestione, la manutenzione e lo sviluppo dei diversi archivi digitali connessi alla galassia Musisque Deoque (www.mqdq.it) e delle funzionalità di indagine ad essi legate [1]. Il progetto è stato condotto entro la cornice e con il supporto del centro VeDPH di Ca' Foscari, Venezia e con la collaborazione del CNR-ILC.

Per quanto riguarda gli aspetti filologico-letterari, obiettivo generale del team di lavoro è un'attività di ricerca scientifica entro il corpus della poesia latina che permetta la ricostruzione della "memoria poetica e del mestiere letterario": MQDQ permette di ricostruire, indagando corpora diversi ma culturalmente inscindibili l'uno dall'altro, una rete fitta di rapporti di dipendenza reciproca, dove la memoria poetica si presenta ora come riuso di espressioni risemantizzate, ora come interiorizzazione inconsciente di una stringa fonica, ora come scelta contrastiva rispetto a un modello letterario, ora come decisione innovativa rispetto a una tradizione testuale. Tale "memoria poetica" è visibile a "colpo d'occhio" nei risultati delle ricerche entro MQDQ (Fig. 1).

14 occorrenze per la chiave: <b>tityre</b> [prima posizione]	
Cercate le alternative grafiche TITYRE, TYTIRE	
VERG. ecl. 1, 1	▶ <b>Tityre</b> , tu patulae recubans sub tegmine fagi
VERG. ecl. 3, 20	▶ <b>Tityre</b> , coge pecus", tu <i>post</i> carecta latebas. 📄
VERG. ecl. 3, 96	▶ <b>Tityre</b> , pascentis a flumine reice capellas:
VERG. ecl. 9, 23	▶ " <b>Tityre</b> , dum redeo (breuis est uia) pasce capellas,
VERG. georg. 4, 566	▶ <b>Tityre</b> , te patulae cecini sub tegmine fagi.
OBTR. Verg. 1, 1	▶ <b>Tityre</b> , si toga calda tibi est, quo tegmine fagi
CALP. SIC. ecl. 3, 19	▶ <b>Tityre</b> , quas dixit, salices pete solus et illinc,
CALP. SIC. ecl. 4, 163	▶ <b>Tityre</b> , rura prius, sed post cantabimus arma."
TER. MAVR. metr. 1675	▶ ' <b>Tityre</b> tu <i>patulae</i> ': concludit syllaba nomen,

Figura 1. Ricerca dell'invocazione "tityre" in prima posizione nel corpus di testi di MQDQ.

Gli archivi di testi [2,3] comprendono ad oggi:

- i. *Poeti d'Italia in lingua latina*, nata nel 1999 per individuare, censire, raccogliere e digitalizzare i componimenti latini in versi prodotti nell'area geografica o in ambienti culturali italiani, entro un periodo compreso all'incirca

fra il 1250 e il 1600: testi spesso trascurati e altrimenti inaccessibili che forniscono però un patrimonio di fondamentale importanza per comprendere i meccanismi di produzione letteraria e di rapporto con “il classico” che informano l’Europa del Pre-Umanesimo, dell’Umanesimo e del Rinascimento. Ad oggi il corpus comprende 316 autori con 841 opere (486.404 versi, 169.549 parole tipo per circa 3.200.000 token).

- ii. *Musisque Deoque*, nato nel 2005, che comprende *tutto* il corpus della poesia latina “antica” a noi nota (276 autori, 642 opere, 343.709 versi, 170.785 parole tipo per circa 2.300.000 token), comprensiva dei *Carmina Latina Epigraphica*. Questi testi sono inoltre provvisti, per circa la metà, di apparati critici, approntati dagli studiosi più qualificati.
- iii. *Hellenica*, un piccolo corpus di testi poetici greci, che è servito per una prima sperimentazione in ambito greco e che auspicabilmente verrà ampliato in futuro.

Il lavoro del team prevede il costante incremento degli archivi (aggiunta di apparati critici a testi che ne siano ancora sprovvisti. Tra gli ultimi lavori di acquisizione riportiamo Columella (libro 10), Lucrezio (libro 1) e sono attualmente in lavorazione Virgilio *Georgiche* e Ausonio. Per quel che concerne le aggiunte di nuovi testi in *Poeti d’Italia* le ultime acquisizioni sono Ludovico Paterno e Antonio Tebaldeo. Inoltre, il lavoro in corso prevede la sistemazione di *corpora* specifici come quello dei *CLE*, la sistemazione della descrizione dei manoscritti testimoni dei testi, lo sviluppo di nuove componenti tecnologiche e infrastrutturali, tra cui la piattaforma di inserimento dei dati, Wiki-MQDQ, rinnovata nel 2022, e l’implementazione delle funzionalità di ricerca avanzate quali ricerche lessicali in testo e apparato, co-occorrenze lessicali, co-occorrenze metrico-verbali e, infine, prove di scansione su versi non solo esametrici. Per quest’ultimo punto in particolare, nel corso del 2022, sono state sviluppate nuove funzionalità per la scansione dei distici elegiaci provenienti da materiale epigrafico. L’impegno costante sugli archivi MQDQ produce un’ingente mole di contributi scientifici che utilizzano gli strumenti e le potenzialità così offerti: si veda, per esempio, il recente volume Manca-Venuti 2021 [1].

## 2. MQDQ: ARCHITETTURA DEL WORKFLOW DISTRIBUITO

Dal punto di vista infrastrutturale, a partire dal 2020, è stato avviato un progetto tra CNR-ILC e DSU dell’Università Ca’ Foscari Venezia finalizzato alla reingegnerizzazione dell’architettura tecnologica che regge il flusso di lavoro orientato alla produzione, rappresentazione, conservazione, indicizzazione e fruizione dei corpora che costituiscono la galassia MQDQ. Il modello architetturale attualmente implementato è stato progettato per soddisfare le esigenze dei collaboratori di Musisque Deoque, ma è progettato con principi di genericità, estensibilità e riusabilità per accogliere in futuro anche i requisiti di iniziative simili, siano essi requisiti funzionali (ad esempio interrogare il corpus per recuperare specifici luoghi del testo), siano essi requisiti non funzionali (ed esempio ottenere la risposta ad un’interrogazione nel minor tempo possibile). In particolare, la Fig. 2 mostra le componenti e il processo sviluppato per la gestione del flusso di lavoro. L’architettura si divide in quattro attività principali: 1) preparazione delle risorse testuali; 2) conservazione del materiale e accesso distribuito; 3) analisi ed indicizzazione dei testi; 4) fruizione ed interrogazione del corpus.

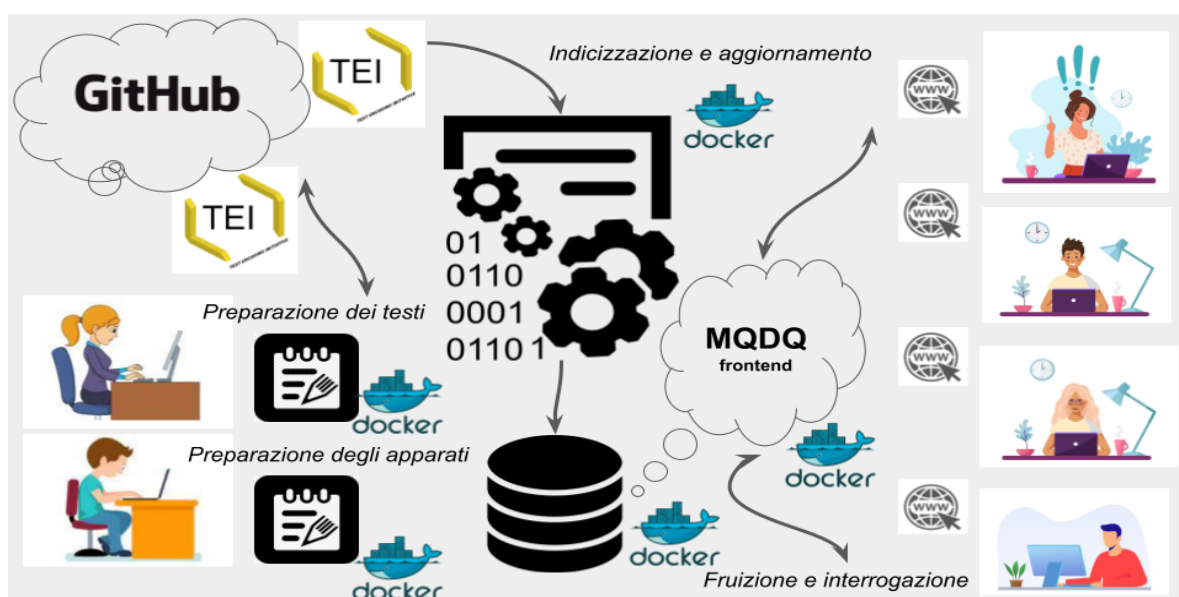


Figura 2. Architettura del flusso di lavoro per la galassia MQDQ.

Per quanto riguarda la fase di preparazione delle risorse, si contano, a loro volta, due distinte procedure. La prima consente di preparare i testi con le opere d'interesse accompagnati da metadati descrittivi; la seconda consente di registrare informazioni d'apparato quali varianti, note a margine, note critiche, etc. collegandoli al testo di base. Le componenti sviluppate fanno uso di tecnologie digitali in grado di velocizzare la compilazione dei materiali testuali. Nello specifico, il testo e i relativi metadati possono essere compilati tanto seguendo lo schema di codifica direttamente in formato XML-TEI [4, 5] (con modello del testo, struttura del documento e schema di codifica personalizzato), quanto mediante un approccio orientato ai Domain-Specific Languages (DSLs) [6]. In Fig.3 è mostrato un esempio di testo annotato mediante approccio DSL con il formalismo derivato dal formato *markdown* (md). E' stato così possibile trascrivere e "marcare" il documento senza particolari competenze tecniche [7]. Grazie al formato *markdown* è semplice assistere lo studioso con *preview* web del testo in fase di editing. Nell'esempio, i versi dell'opera dell'Anfitrione di Plauto sono intuitivamente codificati dal "andata a capo" e le integrazioni presenti nella fonte a stampa sono state annotate per mezzo di una convenzione normata dal formato *md*, con cui è possibile selezionare la porzione di testo interessata con parentesi quadre e quindi associarle una etichetta personalizzata, e.i. *[iam][s]* (dove *iam* è il segmento di testo della fonte e *s* è l'etichetta metatestuale). L'etichetta usata deve essere poi definita in calce al documento attraverso una dichiarazione e una descrizione con testo libero preceduto dal segno cancelletto (*[s]: # 'supplied'*). Infine, il modulo software deputato alla gestione dei testi, tradurrà automaticamente il documento così annotato in formato XML-TEI.

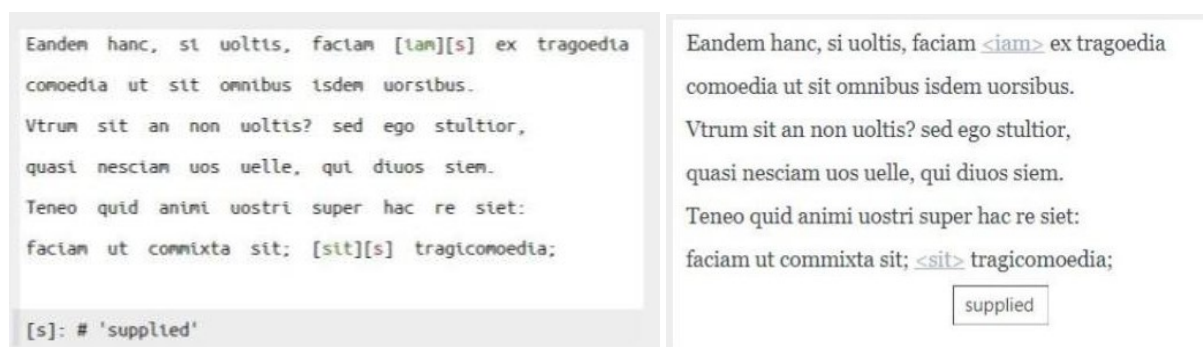


Figura 3. Editing testuale mediante formato di codifica markdown con adattamento preview HTML.

Le informazioni d'apparato<sup>1</sup>, invece, hanno una struttura più complessa ed articolata rispetto al testo dell'opera, dovendo infatti registrare una molteplicità di informazioni a partire dalle fonti manoscritte e a stampa, il riferimento al testo, la registrazione del lemma, delle varianti e degli aspetti descrittivi sulle fonti. Anche per questa fase è possibile operare direttamente compilando il documento in formato XML-TEI, facendo uso del metodo *Double End-Point Attachment (DEPA method)*<sup>2</sup>, che garantisce al tempo stesso massima flessibilità e un'ampia espressività per la rappresentazione delle informazioni d'apparato. Tuttavia è stato sviluppato un agile sistema web per la registrazione delle entrate d'apparato che consente di velocizzare e guidare questa delicata fase di analisi filologica dei testi. In Fig. 4 è possibile cogliere le opzioni offerte dal sistema di gestione dell'apparato: mediante una interfaccia grafica molto amichevole, lo studioso seleziona la porzione di testo relativa alla lezione accolta (lemma dell'apparato) per poi registrare le lezioni varianti e i testimoni con semplici operazioni invocate da menù contestuale.

<sup>1</sup> Per la definizione di un linguaggio specifico di dominio volto alla codifica delle informazioni di apparato sono state ad oggi avanzate varie ipotesi di lavoro. Tuttavia, non è stata rilasciata ancora alcuna implementazione. Per questo motivo ci riserviamo di documentarne gli sviluppi in contributi successivi al presente lavoro.

<sup>2</sup> <https://tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPDE>



Figura 4. Produzione dei dati di apparato nel sistema di backend MQDQ.

La fase di preparazione dei materiali testuali fa uso di un robusto ambiente digitale volto alla conservazione e al tracciamento delle modifiche apportate ai dati in editing; si tratta del sistema distribuito di versionamento *git* [8] e della relativa piattaforma web per la gestione dei *repository git* denominata *github*<sup>3</sup>. L'ambiente *git/github*<sup>4</sup> è usato dal progetto sia come sistema per l'editing collaborativo delle risorse testuali, sia per il confronto delle varie modifiche apportate dai collaboratori, nonché per la conservazione dei dati in lavorazione. Tutti i documenti testuali sono ospitati in formato XML-TEI su appositi repository con accesso privato accessibili solo dai membri del progetto. Le versioni stabili dei corpora sono invece depositati e accessibili mediante l'infrastruttura CLARIN<sup>5</sup>, la quale non solo espone una scheda descrittiva della risorsa, ma rilascia anche un PID handle per l'accesso a lungo termine ai materiali pubblicati per mezzo dell'infrastruttura (<http://hdl.handle.net/20.500.11752/OPEN-555>). Una volta rilasciata la versione pubblica del corpus, i dati testuali corredati delle informazioni di apparato subiscono un processo di analisi e di indicizzazione durante il quale una complessa procedura finalizzata al trattamento automatico del testo elabora i documenti codificati per le scansioni metriche e per la produzione di complessi indici di accesso ed interrogazione al testo. Ad oggi, gli indici sono fisicamente implementati in tabelle relazionali e ottimizzati per ricerche full-text sulle lezioni a testo e in apparato, corredando i dati testuali con informazioni descrittive, bibliografiche e metriche. Terminata la fase di indicizzazione, il corpus è pronto per poter essere fruito dagli studiosi ed interrogato con gli strumenti d'indagine messi a disposizione dall'ambiente web di Musisque Deoque.

### 3. VERSO L'INTERROGAZIONE DEL CORPUS MEDIANTE API REST

Tra gli obiettivi più importanti dell'attuale progetto della galassia MQDQ c'è la proposta di affiancare all'ambiente di fruizione web mediata da interfaccia grafica, anche una modalità di accesso "programmatico" mediante interfaccia API REST [9]. I servizi esposti in rete e raggiungibili tramite chiamate HTTP ricalcano le funzionalità offerte dall'ambiente GUI. Tuttavia la disponibilità offerta dall'interfaccia programmatica estende notevolmente il potenziale d'utilizzo sia per analisi massive dei testi raccolti nel corpus, sia per lo sviluppo di nuovi servizi nonché per l'implementazione di ulteriori applicazioni di fruizione web.

<sup>3</sup> <https://github.com/mqdq-galaxy>

<sup>4</sup> Per la gestione in hosting dei repository *git* si sta tuttavia valutando l'opportunità di migrare dalla piattaforma *GitHub* verso soluzioni completamente aperte quali la piattaforma *GitLab* (<https://about.gitlab.com/company/>).

<sup>5</sup> <https://www.clarin.eu/> - <https://www.clarin-it.it/> - <https://ilc4clarin.ilc.cnr.it/>

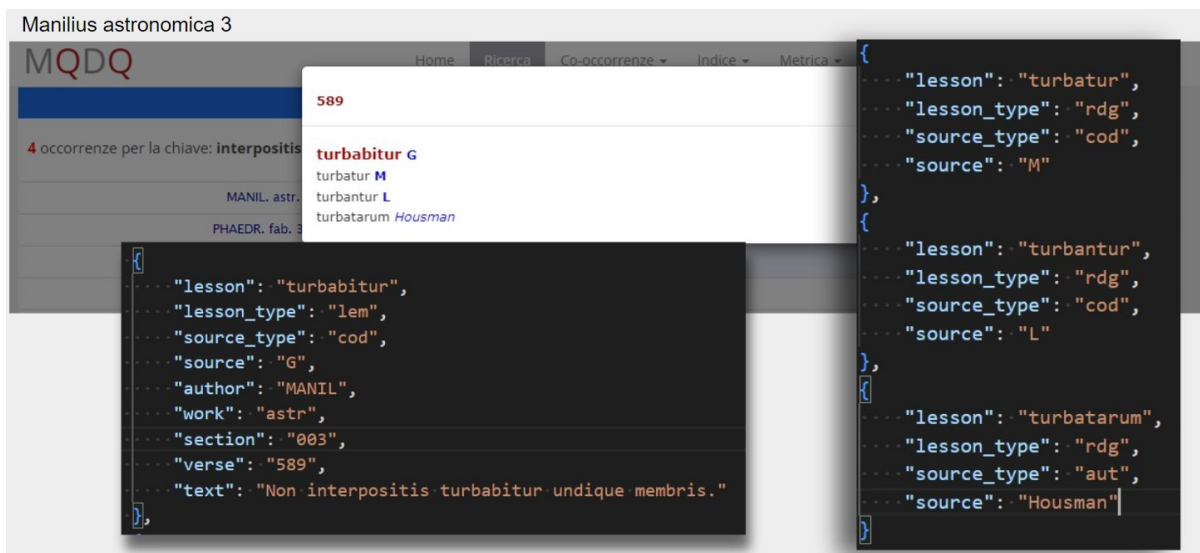


Figura 5. Risultato in formato json per interrogazione tramite API REST.

Ad oggi sono disponibili API espone con metodo GET finalizzate al recupero dei dati gestiti dal sistema. I servizi saranno documentati e pubblicati mediante sistema *OpenAPI/Swagger*<sup>6</sup>. In particolare la versione attualmente sviluppata (v1) espone end-point REST<sup>7</sup> per ottenere

1. la lista degli autori: `http://{BASE-API-URL}/v1/rest/indici/autori`
2. il numero di autori: `http://{BASE-API-URL}/v1/rest/indici/count/autori`
3. i lavori di un autore: `http://{BASE-API-URL}/v1/rest/indici/autori/{idAuthors}`
4. la lista delle opere con metro: `http://{BASE-API-URL}/v1/rest/indici/metri`
5. le sezioni di un'opera mediante codice mqdq (*AUTORE|opera*, ad esempio *VERG|eclo* per le Ecloghe di Virgilio): `http://{BASE-API-URL}/v1/rest/texts/{codice}`
6. la lista di versi di una sezione mediante codice (*AUTORE|opera|sezione*, ad esempio *VERG|eclo|001*): `http://{BASE-API-URL}/v1/rest/texts/{codiceSezione}`
7. i versi di un particolare metro: `http://{BASE-API-URL}/v1/rest/texts/{codiceMetro}`
8. le occorrenze di forme flesse anche mediante *wildcard*: `http://{BASE-API-URL}/v1/rest/texts/ricerca?query={term}`
9. l'autore di una occorrenza specifica di una specifica opera: `http://{BASE-API-URL}/v1/rest/texts/ricerca/autore?codice={codiceTerm}`
10. l'opera di una occorrenza specifica di uno specifico autore: `http://{BASE-API-URL}/v1/rest/texts/ricerca/opera?codice={codiceTerm}`
11. la citazione completa di una singola occorrenza: `http://{BASE-API-URL}/v1/rest/texts/ricerca/citazione?codice={codiceTerm}`
12. le occorrenze di multiword: `http://{BASE-API-URL}/v1/rest/texts/ricerca/multiword?query={}&parole={}&distanza={}&ordine={}`
13. occorrenze varianti: `http://{BASE-API-URL}/v1/rest/texts/search/variant?query={Term}`
14. informazioni d'apparato in forma breve: `http://{BASE-API-URL}/v1/rest/texts/search/lessons?Word={codiceTerm}`
15. informazioni d'apparato completa: `http://{BASE-API-URL}/v1/rest/texts/search/exlessons?Word={codiceTerm}`

La figura 5 mostra un esempio relativo alla chiamata API per la parola *Turbatur* ottenuta con la seguente stringa: `http://{BASE-API-URL}/v1/rest/texts/search/exlessons?Word=TVRBATVR`

<sup>6</sup> <https://swagger.io/specification/>

<sup>7</sup> L'applicazione può essere installata anche in locale, ma si può trovare un endpoint dimostrativo sostituendo {BASE-API-URL} con 48009.omega.ilc.cnr.it, ad esempio: <http://48009.omega.ilc.cnr.it/rest/indici/autori>

## 4. CONCLUSIONI

Nei progetti della Galassia MQDQ le funzionalità descritte in questo contributo e il loro sviluppo futuro devono rispondere a due principi fondamentali, che informano la filosofia degli archivi e rendono MQDQ riconoscibile rispetto ad altri progetti: 1) la semplicità e intuitività di consultazione, che hanno il preciso scopo di generare nell'utente "ispirazione" per le sue stesse ricerche nonché importanti ricadute nell'utilizzo didattico dello strumento, e 2) i tempi di risposta del motore di ricerca che assicurano la presentazione dei risultati dell'indagine con una latenza minima. A questi due principi si è aggiunto, negli ultimi due anni, lo studio di un modello di manutenzione e gestione efficiente, sostenibile ed estremamente raffinato, che può essere applicato a progetti analoghi a MQDQ. Ci si sta già muovendo in questa direzione per il progetto FLOS su Florilegi Siriacei (<https://www.unive.it/pag/40548>) e RiALFRI (<https://www.rialfri.eu>) per la ricerca integrata su testi in poesia e in prosa. Vale la pena infine menzionare lo sforzo che si sta compiendo per integrare il motore di ricerca di MQDQ nel quadro generale della Federated Content Search promossa dall'infrastruttura CLARIN (<https://contentsearch.clarin.eu/>). Per quanto riguarda il latino, CLARIN affianca all'impresa di rendere interoperabili e fruibili via web le risorse digitali di natura linguistica [10] una nuova prospettiva di fruizione tramite API di dati testuali comprensivi della variantistica, che va in parallelo con le campagne di acquisizione dei metadati di autorevoli corpora di testi in versi e in prosa, anche al di fuori della Galassia MQDQ [11].

## BIBLIOGRAFIA

- [1] Manca, Massimo, e Martina Venuti. 2021. *Paulo Maiora Canamus Raccolta Di Studi per Paolo Mastandrea*. Fondazione Università Ca' Foscari.
- [2] Mastandrea Paolo, e Linda Spinazzè (a cura di), *Nuovi archivi e mezzi d'analisi per i testi poetici*, Amsterdam 2011.
- [3] Mastandrea Paolo, e Loriano Zurli (a cura di), *Poesia latina, nuova e-filologia. opportunità per l'editore e per l'interprete*, Roma 2009.
- [4] Pierazzo, Elena. 2015. *Digital Scholarly Editing : Theories, Models and Methods*. Farnham Surrey: Ashgate.
- [5] Lou, Burnard. 2014. *What Is the Text Encoding Initiative?* Encyclopédie Numérique 3. Marseille: OpenEdition Press.
- [6] Boschetti, Federico, and Angelo Mario Del Grosso. 2020. "L'annotazione di testi storico-letterari al tempo dei social media." *Italica Wratislaviensia* 11 (1): 65–99.
- [7] Zenzaro, Simone, Angelo Mario Del Grosso, Federico Boschetti, and Graziano Ranocchia. 2022. "Verso La Definizione Di Criteri per Valutare Soluzioni Di Scholarly Editing Digitale: Il Caso d'uso GreekSchools." In *Culture Digitali. Intersezioni, Filosofia, Arti e Media*. Università del Salento - AIUCD.
- [8] Chacon, Scott, and Ben Straub. 2014. *Pro Git*. Apress.
- [9] Fischer, Frank, Börner, Ingo, Göbel, Mathias, Hecht, Angelika, Kittel, Christopher, Milling, Carsten, & Trilcke, Peer. (2019, July 10). *Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama*. *Digital Humanities 2019: "Complexities"* (DH2019), Utrecht.
- [10] Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P. McCrae, Émilie Pagé-Perron, Marco Passarotti, Salvador Ros Muñoz, Ciprian-Octavian Truica: *When linguistics meets web technologies. Recent advances in modelling linguistic linked data*. *Semantic Web* 13(6): 987-1050 (2022).
- [11] Federico, Boschetti, Del Gratta Riccardo, Monachini Monica, Buzzoni Marina, Monella Paolo, and Roberto Rosselli Del Turco. "Tea for two": the Archive of the Italian Latinity of the Middle Ages Meets the CLARIN Infrastructure." In *Proceedings of CLARIN Annual Conference 2020*, pp. 121-125. Virtual Edition, 2020.



# Memorizzare o ricordare?

## La memoria come fatto emozionale

Lorenzo Demma<sup>1</sup>, Daniele Silvi<sup>2</sup>

<sup>1</sup> Università di Bologna, Italia – lorenzo.demma@studio.unibo.it

<sup>2</sup> Università degli Studi di Roma Tor Vergata, Italia – silvi@lettere.uniroma2.it

### ABSTRACT

Chi di noi oggi si affiderebbe alla propria memoria per ricordare un numero di telefono o un evento imminente? Per praticità, ma anche per efficienza e sicurezza, di norma deleghiamo a una macchina la memorizzazione delle informazioni relative a pratiche che potremmo definire ‘meccaniche’ come rispettare un appuntamento o ricordarci di un compleanno. Tuttavia, c'è il rischio di delegare memorie sempre più importanti, intime e affettive. La nostra memoria non è solo un insieme di dati neutri, ma anche un sistema complesso di emozioni. Quando fissiamo un ricordo o lo richiamiamo alla mente, non avviene solo un trasferimento di dati. In questo poster intendiamo discutere criticamente, insieme all'uditorio e ai colleghi che vorranno intervenire, la nostra posizione critica nei confronti del binomio memoria/memorizzare, proponendo un dibattito vivo a fronte degli stimoli che proporremo attraverso il poster.

### PAROLE CHIAVE

Memoria; Digital Heritage; archivi digitali

## 1. INTRODUZIONE

Quando ricordiamo, ci emozioniamo e quando ci emozioniamo, ricordiamo. È grazie a queste emozioni e ai ricordi che viviamo che siamo spinti all'azione o alla riflessione. Tuttavia, non si tratta di scegliere se affidare o meno parte della nostra memoria a un contenitore di dati. Il punto che vogliamo sollevare è come vivere questa simbiosi tra noi e la macchina, in modo che la nostra memoria non perda i suoi aspetti affettivi. In altre parole, come ri-umanizzare i dati memorizzati digitalmente, dal momento che quando trasferiamo alla macchina un compito di memoria, non trasferiamo anche i ricordi. Se non vogliamo assolutamente perdere questo aspetto, è giunto il momento di riflettere su una operazione che troppo spesso compiamo appunto ‘meccanicamente’.

## 2. METODOLOGIA E CONTENUTI

Nella sua opera, Walter Benjamin affronta il tema della memoria da due diverse angolazioni: dal punto di vista della condizione dell'uomo nella società e nella cultura di massa; e dal punto di vista della comprensione storica, dove il motivo dominante riporta alla necessità del recupero critico di un passato integrale [1].

Benjamin ha una concezione dinamica della memoria, in quanto essa incarna l'idea di trasformazione. Proprio dalla nostra capacità di memoria costruiamo noi stessi; la memoria è la chiave di quella che Benjamin definisce esperienza autentica (o integrale), l'*Erfahrung*[2]. Frutto del processo di sedimentazione di contenuti nella memoria e del loro ritorno come autocoscienza, la cui condizione di esistenza è messa in crisi dai moderni processi di produzione e riproduzione tecnica. La perdita nella società di massa dell'*Erfahrung* è il venir meno della possibilità di congiungere l'individuo al contesto storico. Proprio con i recenti sviluppi delle tecnologie di archiviazione e di riproduzione digitale ci sembra necessario implementare in questa direzione quella che per Benjamin era la dialettica tra esperienza e memoria, partendo da una memoria non più ‘accumulata’ quanto piuttosto costantemente presentificata e disarticolata grazie alle odierne forme di memorizzazione digitale.

Nelle sue *Confessioni*, Sant'Agostino osserva che lo spirito dell'uomo è nei ricordi [3], senza memoria l'individuo perde la sua rete di capacità, di attività e di funzioni concettuali e cognitive; scompare cioè la sua identità. Ma se la memoria è essenziale per l'individuo, lo è anche per la vita intersoggettiva. Nella mitologia greca le Muse sono figlie di Zeus, simbolo del potere, e di Mnemosine, la dea della memoria. Nelle Muse potere e memoria si fondono, e a loro spetta il compito di impedire agli uomini di dimenticare. Sotto questo ultimo aspetto, oggi le nostre muse sono gli spazi digitali permanenti come l'Internet Archive; i profili social dove carichiamo esperienze, riflessioni, foto, video. Sono i servizi di cloud che ci consentono di archiviare ingenti quantità di file costantemente accessibili.



Noi ci poniamo in maniera critica sul fatto che la digitalizzazione possa replicare la memoria. La memoria artificiale, rispetto alla memoria umana (o biologica), quali cambiamenti ha posto in essere durante il processo di memorizzazione? Quali modificazioni hanno apportato questi cambiamenti nell'esperienza dei soggetti?

Nel nostro poster verranno affrontati principalmente i seguenti quattro punti:

1. L'atto del memorizzare e la funzione della memoria sono due attività diverse
2. Il concetto di memoria collettiva e connettiva, secondo il pensiero di McLuhan [4] e De Kerckhove [5]
3. Le evoluzioni delle forme di memorizzazione dall'analogico al digitale
4. Effetti indesiderati: le strategie di memorizzazione atrofizzano la memoria?

## **BIBLIOGRAFIA**

- [1] Grande, Teresa e Olimpia Affuso. M come memoria. La memoria nella teoria sociale. Liguori, 2012.
- [2] Benjamin, Walter. "Di alcuni motivi in Baudelaire" in Angelus Novus. Saggi e frammenti. Einaudi, 1962.
- [3] Agostino. La Trinità. Città Nuova, 1973.
- [4] McLuhan, Marshall e Bruce Powers. The global village: transformations in world life and media in the 21. century. Oxford University Press, 1992
- [5] Susca, Vincenzo e Derrick De Kerckhove. Transpolitica. Nuovi rapporti di potere e di sapere. Apogeo, 2008.
- [6] Ernst, Wolfgang. Digital Memory and The Archive. University of Minnesota, 2013
- [7] Chun, H.K.W. Programmed visions: Software and memory, MIT Press, 2011

# GECA 3.0 – A new tool for cataloguing and enjoying cultural heritage

Roberta Maggi<sup>1</sup>, Tiziana Pasciuto<sup>2</sup>, Martina Mazzoleni<sup>3</sup>, Maria Teresa Artese<sup>4</sup>, Isabella Gagliardi<sup>5</sup>,  
Riccardo Albertoni<sup>6</sup>

<sup>1</sup> CNR-IMATI Genoa, Italy – roberta.maggi@ge.imati.cnr.it

<sup>2</sup> CNR-IMATI Genoa & University of Genoa, Italy – tiziana.pasciuto@ge.imati.cnr.it

<sup>3</sup> San Martino General Hospital, Italy – m.artinamazzoleni@hsanmartino.it

<sup>4</sup> CNR-IMATI Milan, Italy – teresa@mi.imati.cnr.it

<sup>5</sup> CNR-IMATI Milan, Italy – gagliardi@mi.imati.cnr.it

<sup>6</sup> CNR-IMATI Genoa, Italy – riccardo.albertoni@ge.imati.cnr.it

## ABSTRACT

Cataloguing is an essential component of studying and preserving our history. With the development of new systems of digital information, a new panorama where heritage and new technologies coalesce has been forged.

Due to the COVID-19 pandemic, the chance to use and enjoy cultural resources through online platforms has become a true need and has grown exponentially: the concrete impossibility to study, consult and observe “live” any type of material stored in libraries, museums, and archives highlighted the importance of making cultural heritage accessible and available to every type of user.

This paper aims to present and describe GECA 3.0, a platform for integrated management of bibliographical, archival and museum resources and enjoyment of catalogued cultural heritage, characterized by a triple soul: one is particularly centered on cataloguing, another one is a halfway between cataloguing and enjoyment, and the last one purely focused on enjoyment. In order to provide a useful tool not only to catalogue and describe, but also to enjoy and enhance the cultural heritage, a platform to integrate, manage and display data and information from archives, museums, and libraries is being studied and developed.

## KEYWORDS

GECA; GLAM; Cultural Heritage standards; Cataloguing; Enjoyment of Cultural Heritage

## 1. INTRODUCTION

Among the actions for the protection and enjoyment of cultural heritage, cataloguing is an essential component of studying and preserving our history.

Indeed, cataloguing itself should be considered as part and parcel of the events that concern the cultural object ([7]). We must consider the large number of objects that, due to the lack of funds and cataloguers, have never been studied and catalogued: these “not described objects” don't exist.

With the development of new systems of digital information, a new panorama where heritage and new technologies coalesce has been forged ([1]). As reported by Giaccardi and Palen, “the encounter between the complex reality of heritage and ICT is not only an opportunity but also a need” ([4]).

There are many conditions that promote the development of cataloguing systems: digital consultation, for example, allows access to documentation physically preserved on the other side of the world; moreover, it encourages the preservation of cultural objects themselves, especially of the most fragile, whose physical consultation is impossible or strongly advised against.

On the other hand, sometimes navigation in cataloguing systems or online platforms can be challenging for a general user or for professionals who are usually unfamiliar with digital solutions ([18]; [27]). To avoid confining the catalogued cultural heritage in a digital “limbo” known only by researchers and scholars, it is necessary to approach every kind of target audience, for example even the fans and the simply curious ones. This process could be implemented through the development of thematic routes that encourage not only web search, but also an interaction with the cultural object.

This is exemplified in the work undertaken by Nanyang Technological University in Singapore with a project of engineering historical memory and interactive exploration of archival documents ([22]). In this way, digital documents are no longer just a “digital reproduction” that replaces the original for preservation, dissemination, and fruition reasons, but become a tool for deepening and amplifying the meaning and content of the original itself ([3]; [28]).

Due to the COVID-19 pandemic, the chance to use and enjoy cultural resources through online platforms has become a true need and has grown exponentially: the concrete impossibility to study, consult and observe “live” any type of material stored in libraries, museums, and archives highlighted the importance of making cultural heritage accessible and available to every type of user ([17]). This urgency also implies the need to describe and digitize the resources currently available only through consultation, using standardized descriptive criteria as complete and detailed as possible, to guarantee through the study of the “virtual” item the same quality (or superior) provided by the “real” consultation of a cultural object ([26]). Last but not least, the need for action in this GLAM sector was clearly expressed by the recent National Plan for the Digitization of Cultural Heritage ([21]), promoted by the Italian Ministry of Culture in agreement with the Regions. The plan is aimed at cultural institutions that have the task of preserving, protecting, managing and enhancing cultural heritage and aims to promote the process of digital transformation in the five-year 2022-2026, taking into account the objectives of the National Recovery and Resilience Plan (NRRP).

## 2. GECA

GECA<sup>1</sup>, acronym for GEstione CAtaloghi, is a cataloguing system born in 2004 and developed by CNR IMATI. For years, it has been used as a system of management and cataloguing of bibliographical resources of the National Research Council. However, the requirements emerged from ongoing research projects, application scenarios and needs not originally foreseen in GECA, have led to the evolution of the tool, expanding the landscape to other cultural heritage domains, such as archival and museal ones, with the aim of providing the community a useful tool. Thus, this has led to a recent re-engineering of the platform, which is currently being studied and developed.

The tool is now characterized by a triple soul: one is particularly centred on cataloguing, another one is a halfway between cataloguing and enjoyment, and the last one purely focused on enjoyment.

In detail, GECA can be considered, respectively: (i) as a system of cataloguing and describing, respecting its original nature and the cataloguing needs of the National Research Council libraries, ensuring a rigorous bibliographical, archival and museum description, as complete as possible and focused on domain standards; (ii) as source guide: it represents an addressing tool for the user and stands as a source of first access to all the materials available in GECA pertaining to a specific topic or field, which can be identified from the "Project section" linked to each catalographic record. Therefore, the source guide indicates all cultural resources that are part of a project, related to a specific topic, context or historical figure; (iii) as an enjoyment tool, contributing to the creation of thematic routes that provide for the cultural object to be perceived, experienced, and studied. The thematic routes allow the management, the visualization and the fruition of the information coming from the three different domains, with the possibility to integrate them with other information or multimedia materials external to the platform (audio, video, images). Eventually, the various search routes and paths should support a range of experiences and cater for different users' needs and interests as well as personalised ways to relate to the cultural heritage and learn more about it or its context.

## 3. MATERIALS AND METHODS

Before deciding to develop a tool for the integrated management of resources belonging to different cultural heritage domains, an analysis of the main tools available on the market was carried out. The research showed that, despite the availability of integrated tools, most of them do not cover the integrated management of all three domains, are paid tools and at most provide free trial versions with limited time. Also, the characterization of resources as belonging to a source guide and/or a thematic path is not foreseen.

In order to conciliate the three souls and address the requirement stemmed by the new projects, new data models have been developed for the creation of new cataloguing cards in GECA. In particular, an analysis of the main national and international standards and guidelines used in the domain of cultural heritage (libraries, archives, and museums) was carried out. Were analysed the standards created by authoritative bodies in the domain, the most used and recognized by the reference communities (i.e. bibliographic domain: IFLA, Library of Congress; archival domain: ICA, SAA; museum domain: Getty Institute, ICCD).

One of the steps to be addressed to develop the GECA system consisted in comparing and mapping the most used national and international descriptive standards adopted in archives, museums, and libraries. Currently, the bibliography on the realization of a complete, updated and/or freely available mapping among the main descriptive standards is rather sparse or dated ([2]; [16]). For example, in the Tag Library of the EAD3 standard a partial mapping between the ISAD(G) - EAD3 and MARC21 - EAD3 standards is reported. Instead, a mapping between EAD3 and UNIMARC is totally missing, since these are standards of American and European origin respectively ([9]; [24]). Moreover, the conversion sheets from

---

<sup>1</sup> <http://geca.imati.cnr.it>

UNIMARC to MARC21, made available by the Library of Congress, are currently at version 3.0, dated to August 2001 ([20]), despite the standards being updated respectively to the 3rd edition in 2008 ([9]) and Update No. 35 in 2022 ([19]). The purpose of the mapping is to describe the catalogued objects in a lingua franca, allowing the interchange of data between standards. The standards mapping is key to developing new cataloguing data sheets, to ease the insertion of information also for those who are not professionals in the domain (as opposed to archivists, librarians, museum cataloguers) and to support complex thematic routes which need to integrate information coming from separated domains. The adopted descriptive reference standards for the management of cultural bibliographic, archival and museum resources are UNIMARC ([9]), MARC21 ([19]), and ISBD ([8]) for the bibliographic domain, ISAD ([6]), ISAAR ([5]), EAD ([25]) and EAC ([24]) for the archival domain, ICCD ([15]) for the museum domain. In addition, transversely to the domains, an export in MARC21/UNIMARC standard formats wants to be provided, not only to bibliographical resources, but also to archival and museum ones, as has already been done on other cataloguing systems, i.e. the collective database of the Galileo Museum in Florence<sup>2</sup>.

The new data models, inspired by the ICCD cataloguing cards, aims to be all-encompassing, similar (but not identical) for each type of cultural heritage and wants to offer the possibility of viewing data according to the domain standards. In detail, different data models for cataloguing have been developed: (i) four for the bibliographic domain, that are named RB, SP, LG, and RA, respectively to describe bibliographic resources, such as monographs and serials, articles, grey literature, and in access resources, such as the serial resources electronically managed; (ii) two for the archival domain, appointed BAL and BAU, to represent the archival resources from the general to the specific; (iii) three for the museum domain, named PST, OA, and F, centred on the description respectively of scientific and technological heritage, objects, and works of art, and photographs.

As for the bibliographic domain, the data models and the catalogue cards have been developed freely inspired by the model of the ICCD guidelines ([15]). Regarding the archival domain, BAU and BAL cataloguing cards coming from a cross-analysis of the S-Prints ([14]) and F-Photograph ([11]) ICCD guidelines. Finally, as for the museum domain, the models provided by the ICCD guidelines for the PST-Scientific and Technological Heritage ([13]), OA-Objects and Works of Art ([12]), and F-Photograph ([11]) have been directly used. As for intangible cultural heritage, the IAV-Iconographic Heritage and Audio/Video card, inspired by the ICCD's BDI-Intangible Demo-Ethno-Antropological Heritage ([10]), will be soon developed.

Moreover, an additional data model and relative catalogue card, named PPS Storytelling, has been designed. The latter catalogue card was created to enjoy the catalogued cultural heritage. It can have several purposes, which are highlighted by its acronym PPS, which stands for Projects, (Thematic) Paths, Stories. For example, as «Project», it can enclose in a single virtual container all the resources of a given project of cataloguing and description or source guide, while as a "Thematic Paths" can allow the enhancement of the heritage related to a specific project or a specific theme.

Specifically, the Storytelling card, which continues to be studied and its development is ongoing, contains different kinds of information. The PPS card opens with a title and a description of the content; information on the topics dealt and the chronological arc that concerns the heritage enclosed in it are provided. Links to cultural catalogued resources represent the heart of the card: in the middle part they are collected, divided by object type (i.e. books and journals, articles, archival resources, scientific-technological heritage, photographs, and so on). To consult the resource of interest, just click on the arrow next to the title to be redirected to the cultural object's cataloguing card. In addition to links to resources stored within GECA, links to external resources are possible, for example by inserting links pointing to websites and portals. The card can also be enriched with images, video and audio resources.

## 4. RESULTS AND CASE STUDIES

### 4.1. Geca as system of cataloguing and describing

Among the resources catalogued in GECA over the years, a remarkable *corpus* of ancient books has been catalogued. The collection, concerning the history of technology, consists of about a hundred books published between 1520 and 1830.

These bibliographical cards are among the most complete and rich within GECA: for each of the resources, has been made not only the specimen description, but also have been inserted historical, biographical, artistic, cultural, and social information, in order to promote and enhance the National Research Council ancient book heritage (Figure 1).

---

<sup>2</sup> [https://opac.museogalileo.it/imss/search?collection\\_data=0&s=25&o=score&q=&v=l&h=any\\_bc&l=en](https://opac.museogalileo.it/imss/search?collection_data=0&s=25&o=score&q=&v=l&h=any_bc&l=en)

To not disperse all these important resources within the GECA platform, the Storytelling card allowed to gather together in a thematic path, which was also enriched by a description on the importance of the *corpus* and from attached images<sup>3</sup> (Figure 2)

**ANCIENT BOOK, PRINTED**

**Gli artiftios et curiosi moti spiritali di Herrone. Tradotti da M.Gio. Battista Aleotti d'Argenta. Aggiuntoui dal medesimo quattro theoremi non men belli, & curiosi de'gli altrj. Et il modo con che si fa artificiofamente salir vn canale d'acqua viua, ò morta, in cima d'ogn'alta torre. Al serenif.mo Signore D. Alfonso II. Duca di Ferrara suo signore**

**Author** Heron, Alexandrinus

**Contributor** Aleotti, Giovanni Battista (Translator); Baldini, Vittorio (Printer)

**Publication** In Ferrara : per Vittorio Baldini, Stampator Ducale, 1589

**Physical description** [12], 103 p., [1] : ill. ; 4\*

**Identifier** Fingerprint uouon he69 K.o, frgu (3) 1589 (R)

**Country** Italia

**Language** Italiano (The document is the translation of the original work or an intermediate work)

**Contained in**

**Thematic route** I libri antichi del Centro di Studio sulla Storia della Tecnica del CNR di Genova

**Thematic route** I libri antichi del CSST del CNR di Genova che trattano di energia

**SUMMARY OR ABSTRACT**

**NOTES**

**EDITORIAL MARK**

**HELD BY**

IT/IRRC/00308645 Armored Last update 02/05/2023



See original by BYTERFLY



Main image (Cover)

**SHOW BY**

LABELS ISBD

BROWSE

UNIMARC XML UNIMARC

MARC21 XML MARC21

Figure 1. Cataloguing card of an ancient book in GECA.

**THEMATIC ROUTE**

**The ancient books of the Study Center on the History of Technology of the CNR in Genoa**

**The bibliographic heritage**

The collection was established from the late '60s by Carlo Maccagni (1932 – 2022), professor of History of Science at the University of Genoa and director, until 2002, of the Study Center on the History of Technology (CSST, now IRRCES) of the CNR in Genoa.

During his tenure, he created one of the Italian scientific library collections of high specialization, through the acquisition of valuable specimens of antiquarian books, precious anastatic copies of unavailable volumes and collections of scholars, which constitute an important collection of primary sources of technical and scientific topics published over a period of four centuries from the fifteenth to the nineteenth.

The collection includes technical manuals and scientific treatises that cover multiple disciplines: astronomy, geography, physics, mathematics, earth sciences, metallurgy and steel industry, artillery and ballistics, mechanical and hydraulic engineering, architecture; there are also many car theaters.

The collection contains several manuscripts and large encyclopedic works, accompanied by tables of technical drawings.

In particular, a substantial portion of the heritage is represented by treatises on mathematics, physics and terrestrial and astronomical geography. Among the latter are numerous treatises on the art of navigation often combined with studies of chronology, as well as technical–theoretical treatises for the realization of clocks and other apparatus for measuring time (gnomons, hourglasses, astrolabes, etc.).

There are also many works dedicated to earth science and techniques related to mining, metallurgy and pre-industrial manufacturing for both civil and military purposes.

The number of volumes dealing with mechanical and hydraulic engineering is still high: many texts that theorize and describe construction techniques of hydraulic and defence works, walls; and a substantial corpus of texts includes treatises on the construction and functionality of work machines and especially war. Among the sources of mechanical argument are volumes that transfer skills in systems for the generation and use of energy.

There are also some treatises on the cultivation of tree species and several volumes on techniques and agricultural machinery.

The collection also includes encyclopedias on techniques and others more focused on specific subjects.

A working group composed of researchers and technicians of the Institute of Research on Sustainable Economic Growth, Grazia Biorci, Giancarlo Birello and Anna Perin, thanks to various funding obtained by national calls for the digitization and reuse of valuable bibliographical material, has carried out valorization projects such as permanent exhibitions and multimedia installations online and in museum sites.

**Da ... Mundus subterraneus ... 1678**

**Da Libro dabacco che insegna ... [1625]**

**Is part of**

**Thematic Route** I libri antichi del CSST del CNR di Genova che trattano di energia

**Time period** 1520 - 1830

**Included Objects**

**BOOKS AND JOURNALS:** 61

**STORYTELLING:** 1

Figure 2. Thematic path in GECA, which collects the ancient books of the Study Center on the History of Technology of the CNR in Genoa.

<sup>3</sup> <http://geca.imati.cnr.it/record.php?idk=IT/IRRC/PPS/0000011>

## 4.2. Geca as source guide and enjoyment tool - the digitxl project

The DigitXL Project ([23]), promoted by the National Academy of Sciences called the XL<sup>4</sup> and the Department of Social Sciences and Humanities, Cultural Heritage of the National Research Council (CNR-DSU), consists of a census of the scientific writings produced by the academic members during the history of the Academy, in their bibliographical description, digitalization with criteria and modalities based on advanced technological standards to enable consultation on the web.

For the project, GECA represents the management tool of the online digital edition of the works of the members of the National Academy of Sciences, starting from the initial nucleus of 40 scientists who founded the association in 1782. The initiative starts with the development of the pilot project “Novum Corpus Fontanianum”, focused on the figure of the scientist and naturalist Felice Fontana, one of the first members of the Academy of Sciences called XL, as well as founder and first director of the Galileo Museum in Florence. Cultural resources, physically preserved in different international conservation institutes, were described in GECA in simplified catalogue cards, with essential information for the description of the resource, with homogenous criteria and according to updated standards, managed by the upper mentioned platform and shared on web.

Moreover, the Storytelling card was fundamental for the project, in order to (i) aggregate all the resources of the DigitXL Project, whose data have been provided by numerous international institutions, in a single virtual container; (ii) enhancing the heritage relating to the figure of Felice Fontana, through the implementation of specific thematic paths. All the cards made for the Fontana project, therefore, have been collected in a project called “Novum Corpus Fontanianum. DigitXL Project - Felice Fontana”<sup>5</sup> (Figure 3). At the same time, with the aim of recounting the figure of Felice Fontana and describing, more generally, the DigitXL project, six thematic routes were suggested ([23]).

**PROJECT** HOME / BACK

### Novum Corpus Fontanianum

DigitXL Project – Felice Fontana  
The project collects all the resources related to the studies and experiments of Felice Fontana carried out during his life.

**Subject**

**Name** Felice Fontana  
**Title** Physicist, anatomist, biologist  
**Nationality** Italia  
**Was born in** Pomarolo (Trento) the 15 april 1730  
**Dies in** Florence the 19 march 1805  
**Training** He studied philosophy, physics and anatomy in various Italian cities. Fontana studied anatomy and physiology at the University of Padua. In 1765 he was appointed professor of logic in Pisa; the following year he was called by the Grand Duke of Tuscany Pietro Leopoldo di Lorena in Florence and appointed court physicist.  
**Profession** He was an Italian physicist, anatomist and biologist. He was also a distinguished toxicologist and scholar of the human eye.

**Time period** 1760 - 1806

**BIOGRAPHY**

**Included Objects**

**BOOKS AND JOURNALS:** 66  
**ARTICLES:** 97  
**SCIENTIFIC-TECHNOLOGICAL HERITAGE:** 27  
**ARCHIVAL LEVELS:** 11  
**ARCHIVAL ITEMS:** 519  
**STORYTELLING:** 6

Portrait of Felice Fontana. ©Bianchi G., Chelazzi G., Il Museo di Storia Naturale dell'Università degli Studi di Firenze, vol. 1 Le collezioni della Specola: zoologia e cere anatomiche, Firenze University Press, 2009.

Bust of Felice Fontana. © La Specola, Florence. From Bianchi G., Chelazzi G., Il Museo di Storia Naturale dell'Università degli Studi di Firenze, vol. 1 Le collezioni della Specola: zoologia e cere anatomiche, Firenze University Press, 2009.

Figure 3. Project "Novum Corpus Fontanianum. DigitXL Project - Felice Fontana" in GECA, which contains both the cultural resources collected and the thematic routes realized.

## 5. CONCLUSIONS AND FUTURE WORKS

The GECA platform, with his triple soul, represents a useful tool to catalogue and describe, to enjoy and enhance the catalogued cultural heritage. Future developments include to provide a visualization of data from different points of view,

<sup>4</sup> <https://accademiaxl.it>

<sup>5</sup> <http://geca.imati.cnr.it/record.php?idk=IT/ItRC/PPS/00000003>



also according to the domain standards, ensure the export and import of data related to bibliographic, archival and museum resources and to expose data in Linked Open Data.

The GECA potential will be further evaluated and tested in new research projects, i.e. the creation of a Portal of the Sources for the History of the Italian Republic, still ongoing.

In addition, the strengths and weaknesses of the already structured thematic routes will be assessed, according to the involvement of different types of users.

## 6. ACKNOWLEDGEMENTS

We would like to thank the Academy of Sciences called the XL, Fondazione Caritro, Cassa di risparmio di Trento e Rovereto and Maurizio Gentilini with which we collaborated for the DigitXL Project.

## REFERENCES

- [1] Alonso-Jiménez, Roberto Francisco, Mar Loren-Méndez, Daniel Pinzón-Ayala, and Francisco Ollero-Lobato. «Heritage Cataloguing in History: Conceptual and Graphical Foundations of Immovable Cultural Heritage Data Bases in the Case of Spain». *Sustainability* 13.19 (2021): 11043.
- [2] Assumpção, Fabrício Silva, and Plácida Leopoldina Ventura Amorim da Costa Santos. «Representation in the bibliographic domain: A look at MARC 21 Formats». *Perspectivas em Ciencia da Informacao* 20.6 (2015): 54-74.
- [3] Campagnolo, Alberto. *Book conservation and digitization: The challenges of dialogue and collaboration*. Project Muse, 2020.
- [4] Giaccardi, Elisa, and Leysia Palen. «The social production of heritage through cross-media interaction: making place for place-making». *International Journal of Heritage Studies* 14.3 (2008): 281-297.
- [5] ICA - Committee on Descriptive Standards. ISAAR (CPF): International Standard Archival Authority Record For Corporate Bodies, Persons and Families. Second Edition, 2003. [https://www.ica.org/sites/default/files/CBPS\\_Guidelines\\_ISAAR\\_Second-edition\\_EN.pdf](https://www.ica.org/sites/default/files/CBPS_Guidelines_ISAAR_Second-edition_EN.pdf)
- [6] ISAD(G): General International Standard for Archival Description, Second Edition, 2000. [https://www.ica.org/sites/default/files/CBPS\\_2000\\_Guidelines\\_ISAD%28G%29\\_Second-edition\\_EN.pdf](https://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_EN.pdf)
- [7] Harvey, David Charles. «Heritage Pasts and Heritage Presents: temporality, meaning and the scope of heritage studies». *International Journal of Heritage Studies* 7:4 (2001): 319-338.
- [8] IFLA. ISBD International Standard Bibliographic Description, 2011. [https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/isbd/isbd-cons\\_20110321.pdf](https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/isbd/isbd-cons_20110321.pdf)
- [9] IFLA UNIMARC Strategic Programme, and Alan Hopkinson. *UNIMARC manual : Bibliographic format*. 3rd ed. De Gruyter Saur, 2008.
- [10] Istituto Centrale per il Catalogo e la Documentazione. Normativa BDI - BENI DEMOETNOANTROPOLOGICI IMMATERIALI, versione 4.00, 2017. <http://www.iccd.beniculturali.it/getFile.php?id=5734>
- [11] Normativa F - FOTOGRAFIA, versione 4.00, 2016. <http://www.iccd.beniculturali.it/getFile.php?id=8144>
- [12] Normativa OA - OPERE E OGGETTI D'ARTE, versione 3.00, 2018. <http://www.iccd.beniculturali.it/getFile.php?id=7508>.
- [13] Normativa PST - PATRIMONIO SCIENTIFICO E TECNOLOGICO, versione 4.00, 2018. <http://www.iccd.beniculturali.it/getFile.php?id=6624>
- [14] Normativa S - STAMPE, versione 3.00, 2018. <http://www.iccd.beniculturali.it/getFile.php?id=8145>
- [15] Normative, 2023. <http://www.iccd.beniculturali.it/it/normative>
- [16] Jakšić, Mirjana. «Mapping of bibliographical standards into XML». *Software – Practice and Experience* 34.11 (2004): 1051-1064.
- [17] Kolokytha, Olga, and Krisztina Rozgonyi. «From social distancing to digital un-distancing: The COVID-19 pandemic and new challenges for digital policy in the cultural and audio-visual sectors». *Journal of Digital Media & Policy* 12.1 (2021): 177.
- [18] Kutzner, Kristin, Thorsten Schoormann, Claudia Roßkopf, and Ralf Knackstedt. «Designing Online Platforms for Cultural Participation and Education: A Taxonomic Approach» In: Ahlemann, Frederik, Reinhard Schütte, and Stefan Stieglitz (eds). *Innovation Through Information Systems*. WI 2021. *Lecture Notes in Information Systems and Organisation* 46 (2021): Cham: Springer.
- [19] Library of Congress. MARC 21 Format for Bibliographic Data. 35th update, 2022. <https://www.loc.gov/marc/bibliographic>
- [20] UNIMARC to MARC21 Conversion Specifications, Version 3.0, 2001. <https://www.loc.gov/marc/unimarc21.html>
- [21] Ministero della Cultura. Piano Nazionale di Digitalizzazione del Patrimonio Culturale 2022-2023, versione 1.1, 2023. [https://digitallibrary.cultura.gov.it/wp-content/uploads/2023/01/PND\\_versione1\\_1\\_gen2023.pdf](https://digitallibrary.cultura.gov.it/wp-content/uploads/2023/01/PND_versione1_1_gen2023.pdf)
- [22] Nanetti, Andrea, and Davide Benvenuti. «Engineering Historical Memory and the Interactive Exploration of Archival Documents: The Online Application for Pope Gregory X's Privilege for the Monastic Community of Mount Sinai (1274) as a Prototype». *Umanistica Digitale* 10.1(2021): 325-357.
- [23] Pasciuto, Tiziana, Riccardo Albertoni, Roberta Maggi, Maria Teresa Artese, Isabella Gagliardi, and Maurizio Gentilini. «Travelling Culture: Define, Implement, Enrich and Disseminate the Digital Cultural Heritage. The “DigitXL Project” Case Study». *Towards Smart and Inclusive Learning Ecosystem, EDEN Research Workshop Proceedings* (2022): 134-139.



- [24] Society of American Archivists. Encoded Archival Context for Corporate Bodies, Persons, and Families Schemata and Tag Library Version EAC-CPF 2.0, 2022. <https://eac.staatsbibliothek-berlin.de/schemata-and-tag-library>
- [25] Encoded Archival Description Tag Library Version EAD3 1.1.1, 2019. <https://www.loc.gov/ead/EAD3taglib/EAD3-TL-eng.html>
- [26] Sweetser, Michelle, and Alexandra A. A. Orchard. «Are We Coming Together? the Archival Descriptive Landscape and the Roles of Archivist and Cataloger». *American Archivist* 82.2 (2019): 331-380.
- [27] Tim, Yenni, Shan L. Pan, and Tao Hua Ouyang. «Museum in the Age of Digital Transformation». *PACIS 2018 Proceedings* 102 (2018).
- [28] Turbanti, Simona. «Recensione: A. Campagnolo. 2020. Book conservation and digitalization: the challenges of dialogue and collaboration, ARC Humanities Press». *Umanistica Digitale* 10.1 (2021): 417-421.

# Libri, memoria e territorio: la digital library di Carlo Alberto

Anna Maria Marras<sup>1</sup>, Andrea Bolioli<sup>2</sup>

<sup>1</sup> Università di Torino, Italia – annamaria.marras@unito.it

<sup>2</sup> Ricercatore indipendente, Italia – andrea.bolioli@gmail.com

## ABSTRACT

In questo articolo presentiamo le principali metodologie adottate per realizzare la *digital library* della “Biblioteca di Carlo Alberto”, composta da una selezione di 60 volumi della biblioteca privata del Castello Reale di Racconigi e da un catalogo online di circa 5100 volumi, organizzati in un sito web multilingue rivolto al grande pubblico e progettato come strumento di collegamento con i percorsi di visita al Castello e con il territorio circostante.

## PAROLE CHIAVE

Digital Library; Cultural Heritage; Text mining

## 1. SINTESI DEL PROGETTO

Tra il 2017 e il 2018 è stato realizzato un progetto di restauro e riqualificazione della biblioteca di Carlo Alberto di Savoia del Castello Reale di Racconigi, finanziato dalla Fondazione Compagnia di San Paolo, dal titolo “Il recupero della Biblioteca carloalbertina, luogo d’incontro tra vita pubblica e privata del sovrano”, ideato e realizzato dall’Associazione Le Terre dei Savoia, in collaborazione con il Castello di Racconigi (Direzione Regionale Musei Piemonte). Oltre al restauro della biblioteca fisica, il progetto ha incluso la creazione di una piccola *digital library* denominata la “Biblioteca di Carlo Alberto”, il cui sito web è consultabile all’indirizzo <https://www.bibliocaralberto.it/>.

La *digital library* è stata progettata non solo come luogo per conservare ed esplorare i contenuti, ma anche come strumento di collegamento con i percorsi di visita al Castello e con il territorio circostante. Nel progetto è stata effettuata la digitalizzazione di una selezione di 60 volumi, consultabili via *tablet* e PC presso la nuova sala di lettura del Castello. Un ulteriore sottoinsieme è stato pubblicato online. Si è trattato del primo passo di un più ampio processo di digitalizzazione che potrà svilupparsi in relazione al successo dell’iniziativa e alla sostenibilità dei costi gestionali, implementabile in fasi successive. Nel progetto sono state svolte le seguenti attività principali:

- catalogazione dei volumi della biblioteca e creazione del catalogo online;
- selezione, digitalizzazione e OCR di un sottoinsieme di volumi;
- creazione del sito web in 3 lingue, definizione di percorsi tematici di lettura, creazione di mappe interattive di luoghi del territorio connessi ai libri;
- indicizzazione del contenuto dei libri e creazione del motore di ricerca;
- pubblicazione delle edizioni digitali su Internet Archive.

Una descrizione ampia del progetto si trova in [1]. In questo articolo riassumiamo brevemente alcuni aspetti fondamentali. La catalogazione dei libri è stata effettuata secondo gli standard di SBN ([2]), partendo da un precedente schedatura di precatalogo effettuata nel 2007-’08, costituita da circa 4900 record. Si è effettuato un lavoro di pulizia e normalizzazione dei dati, comprendente la normalizzazione dei nomi degli autori secondo gli standard SBN e/o VIAF ([3]). Sono state create nuove schede per i volumi della biblioteca assenti nel precatalogo, arrivando ad un totale di 5108 schede bibliografiche. Il catalogo è consultabile nel sito web tramite un motore di ricerca a faccette ed è stata creata una versione scaricabile in formato RDF Turtle [4], secondo gli standard dei *Linked Open Data* bibliografici, utilizzando i vocabolari controllati di Schema.org e di Bibframe.

È stata effettuata una indicizzazione semantica delle opere utilizzando il Nuovo Soggettario ([5]) e altre risorse per la metadazione, in linea con gli scopi di accessibilità e internazionalizzazione del progetto. Tutti i metadati sono stati tradotti in francese e inglese seguendo le indicazioni di SBN (se presenti), i Library of Congress Subject Headings ([6]), i vocabolari dei *linked data* della Bibliothèque nationale de France ([7]).

Portata a termine la catalogazione, sono stati scelti 60 volumi tra quelli ritenuti più interessanti per la creazione dei percorsi di lettura tematici. I volumi sono stati digitalizzati in loco ad alta risoluzione con scanner planetario Atiz BookDrive. Sono

state digitalizzate 15.000 pagine in formato TIFF; di tutte le pagine scansionate è stata effettuata l'OCR semiautomatica e sono state create le copie in JPEG a 300 dpi. Sono stati quindi creati dei file PDF multipagina a colori con testo ricercabile. Per progettare il sito web si è partiti dallo schema di progettazione della *Web Strategy* elaborato dal gruppo *Digital Cultural Heritage* di ICOM Italia ([8]). Lo scopo del sito è far scoprire non solo i libri della biblioteca, ma anche alcuni luoghi di cui parlano i libri o con cui, attraverso le storie e i personaggi, si creano connessioni con il territorio locale, visualizzate attraverso delle mappe interattive.

I 60 libri digitalizzati sono stati organizzati in 8 percorsi di lettura tematici: Carlo Alberto, Casa Savoia, Infanzia, Viaggi e luoghi, Satira e Costume, Acqua e Seta, Cucina e Piante, Lingua Piemontese.

Il progetto si inserisce in quel filone delle *narratives Digital Library* ([9]) che, partendo da dati strutturati e di catalogo e sfruttando le tecnologie digitali, consentono di creare nuovi racconti e di esplorare connessioni tra i contenuti in grado di coinvolgere le comunità e di farsi strumenti di memoria.

È stata effettuata l'indicizzazione dei testi dei libri e l'estrazione automatica delle parole chiave, con correzione manuale. Nelle pagine di descrizione dei singoli libri nel sito web sono visibili le *cloud* dei termini estratti, oltre alle descrizioni dei libri (redatte appositamente nel progetto) e ai principali metadati. Nel motore di ricerca è possibile effettuare ricerche dentro i testi delle opere.

Per la pubblicazione online di una selezione di 20 dei 60 libri è stata utilizzata la *digital library* Internet Archive ([10]). La scelta di condividere i volumi in Internet Archive è stata dettata dalla volontà di condividere con un vasto pubblico una parte dei volumi, dare maggiore visibilità ai libri stessi e alla biblioteca del Castello di Racconigi, ridurre i costi di gestione e la sostenibilità della *digital library* di Carlo Alberto.

## BIBLIOGRAFIA

- [1] Marras, Anna Maria. «Storytelling e Digital library». *Biblioteche oggi* 41, no. 2 (2023): 36-42
- [2] ICCU Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche. “Normative catalografiche. Regole Italiane di Catalogazione.” <https://www.iccu.sbn.it/it/normative-standard/normative-catalografiche/>
- [3] VIAF Virtual International Authority File <https://viaf.org/>
- [4] W3C Recommendation, “RDF 1.1 Turtle. Terse RDF Triple Language” 25 February 2014 <https://www.w3.org/TR/turtle/>
- [5] Biblioteca Nazionale Centrale di Firenze. “Nuovo Soggettario”. <https://thes.bncf.firenze.sbn.it/>
- [6] The Library of Congress. “The Library of Congress Subject Headings (LCSH)”. <https://www.loc.gov/aba/cataloging/subject/>
- [7] Bibliothèque nationale de France. “Vocabulaires employés à la Bibliothèque nationale de France”. <https://data.bnf.fr/vocabulary>
- [8] Marras, Anna Maria. “Accessibilità digitale”, in Sarah Dominique Orlandi, Gianfranco Calandra, Vincenza Ferrara et al., *Web Strategy museale*, ICOM Italia, 2019, pp. 40-51
- [9] *Narratives in digital libraries* <https://dlnarratives.eu/project.html>
- [10] Internet Archive: <https://archive.org>

# Un modello di catalogazione digitale delle grammatiche bilingue di spagnolo per italofoeni

Silvia Muzzupappa<sup>1</sup>, Giorgio Maria Di Nunzio<sup>2</sup>

<sup>1</sup> Università di Padova, Italia – silvia.muzzupappa@studenti.unipd.it

<sup>2</sup> Università di Padova, Italia – giorgiomaria.dinunzion@unipd.it

## ABSTRACT

Il progetto rientra nel campo di studio della grammaticografia della lingua spagnola come L2 (le cosiddette *Gramáticas de español para italianos*, che d'ora in poi chiameremo semplicemente GEI), in particolare relativo alla produzione grammaticale per l'insegnamento dello spagnolo ad un pubblico italofono nell'arco di tempo compreso tra il XVI ed il XIX secolo. L'obiettivo è stabilire la definizione e i limiti del genere testuale grammaticografico delle GEI utilizzando un database per la raccolta, la catalogazione e l'elaborazione dei dati e delle informazioni estratti dai documenti corpus con il fine di facilitarne attraverso un'interfaccia grafica il riconoscimento delle caratteristiche comuni e non comuni.

## PAROLE CHIAVE

Grammaticografia; digital humanities; database; grammatiche bilingue; lingua spagnola; lingua italiana

## 1. INTRODUZIONE

La presente ricerca ha come oggetto di studio le grammatiche spagnole (o ogni tipo di documento con finalità didattica) per un pubblico di lingua italiana pubblicate tra il XVI e il XIX secolo. Questo progetto nasce dal progetto *Epigrama*, sviluppato dall'Università di Padova, che si occupa di creare una biblioteca digitale contenente i documenti che trattano dell'insegnamento della lingua spagnola in Italia nel corso dei secoli<sup>1</sup>. Infatti, all'interno di *Epigrama* si digitalizzano questo tipo di testi per renderli accessibili ad esperti linguistici e ricercatori del settore e consentire loro di navigare al loro interno. Inoltre, ci sono pochissimi studi focalizzati sull'evoluzione di questo genere specifico attraverso i secoli<sup>2</sup>. Troviamo invece numerosi studi e riferimenti sulla produzione grammaticale sia italiana<sup>3</sup> che spagnola<sup>4</sup> prese singolarmente, non da un punto di vista contrastivo tra le due lingue.

Quindi, a questo punto le domande di ricerca sono:

1. Le grammatiche spagnole per gli italofoeni sono un genere autonomo?
2. Se sì, qual è la definizione di questo genere?
3. E quindi, come individuare i tratti tipici di questo tipo di grammatiche?

Da queste domande si è iniziato a progettare un database per catalogare e raccogliere tutti i dati rilevanti dal corpus ampliato a partire da quello di *Epigrama*, mantenendo la relazione esistente tra loro e confrontandoli attraverso tutti i libri. In questo modo possiamo identificare elementi comuni o non comuni di questo tipo di grammatiche e, infine, suggerire una definizione.

## 2. IL CORPUS E L'ANALISI DEI DOCUMENTI

Il corpus iniziale di *Epigrama* è stato arricchito con altre grammatiche raggiungendo un totale di nove grammatiche scritte da nove autori diversi. Si sono selezionati i documenti per il corpus seguendo quattro criteri:

---

<sup>1</sup> Castillo Peña, «Epigrama: Un portal para la edición digital de textos gramaticales».

<sup>2</sup> Gallina, *Prime grammatiche spagnole ad uso degli Italiani (sec. 16.)*; San Vicente e Lombardini, «Gramáticas de español para italofoenos de los siglos XVI-XVIII: elementos lexicograficos y componentes técnico-científicos»; Kukenheim, *Contributions a l'histoire de la grammaire italienne, espagnole et française a l'époque de la Renaissance*.

<sup>3</sup> Fornara, *La trasformazione della tradizione nelle prime grammatiche italiane (1440-1555)*; Trabalza, *Storia della grammatica italiana*.

<sup>4</sup> Calero Vaquera, *Historia de la gramática española: 1847-1920 de A. Bello a R. Lenz*; Gómez Asencio, *El castellano y su codificación gramatical. De 1492 (A. de Nebrija) a 1611 (John Sanford)*; Gómez Asencio, *El castellano y su codificación gramatical. De (B. Jiménez Patón) 1614 a 1697 F. Sobrino*; Gómez Asencio, *El castellano y su codificación gramatical. De 1700 a 1835*; Ramajo Caño, *Las gramáticas de la lengua castellana desde Nebrija a Correas*.

1. Furono pubblicati tra il XVI e il XIX secolo,
2. Sono testi multilingue,
3. Il metalinguaggio è l'italiano e la lingua oggetto è lo spagnolo, ovvero, l'autore usa la lingua italiana per insegnare e descrivere lo spagnolo,
4. Hanno uno scopo didattico, quindi possono anche non essere mere grammatiche.

Il corpus ad oggi comprende, quindi, i seguenti testi:

1. Mario Alessandri d'Urbino, *Il paragone della lingua toscana et castigliana*, Napoli, Mattia Cancer, 1560;
2. Mattia Chirchmair, *Gramatica spagnuola e italiana*, Firenze, Giuseppe Manni, 1709;
3. Antonio Fabro, *Grammatica per imparare le lingue italiana, francese e spagnola*. Venezia, Gio. Dini, 1627;
4. Lorenzo Franciosini, *Gramatica spagnola e italiana*, Venezia, Sarzina, 1624;
5. José Martínez de Valdepeñas, *Grammatica della lingua spagnola*. Genova, Stamperia di Giovanni Franchelli, 1786;
6. Giovanni Miranda, *Osservationi della lingua castigliana*, Vinegia, Gabriel Giolito de Ferrari, 1566
7. Josef Faustino Perles y Campos, *Gramatica Española, o' modo de entender, leier, y escribir Español*, Napoli, Parrino & Mutii, 1689;
8. Gennaro Sisti, *Traduzione dal francese del nuovo metodo di Porto Reale. Con cui agevolmente s'insegna la lingua spagnola*, Napoli, Serafino Porsile Regio Stampatore, 1742;
9. Massimo Troiano, *Compendio del signor Massimo Troiano con le annotazioni del signor Argisto Giuffredi*, Firenze, 1601.

Per riprodurre esattamente la struttura e il contenuto di ogni grammatica si è deciso di lavorare su due livelli diversi ma paralleli: la macrostruttura e la microstruttura.

La macrostruttura concerne l'impaginazione dell'intero libro, nonché la suddivisione del libro in capitoli, sezioni, sottosezioni e paragrafi.

La microstruttura coincide con il contenuto della grammatica stessa, quindi si concentra sulla terminologia utilizzata per la definizione, la spiegazione e la rappresentazione delle parti del discorso di entrambe le lingue trattate nel libro.

### 3. IL DATABASE E LA PROPOSTA DI UNA DEMO

L'utente potrà interagire col database attraverso un'interfaccia grafica, raggiungibile via web, che gli consentirà di applicare filtri e selezioni attraverso una procedura parzialmente guidata. Per il momento il database è in fase di progettazione e implementazione.

I primi filtri richiesti all'utente hanno lo scopo di limitare la ricerca ad un sotto-insieme dei testi, possibilità che può anche essere esclusa dall'utente stesso. Infatti, si tratta comunque di un passaggio opzionale, che sicuramente tornerà utile quando il database assumerà delle dimensioni di molto maggiori, ovvero si amplierà il corpus anche con altri documenti bilingue o multilingue con finalità didattica. Tali filtri permettono quindi di selezionare solo determinati testi, per esempio in base alla lingua usata dall'autore (il metalinguaggio) o alla lingua insegnata (la lingua oggetto), oppure di selezionare le opere appartenenti ad un preciso periodo storico. I valori di filtro per questa fase sono:

- l'autore o gli autori (con la possibilità di selezionarli tutti);
- il libro o i libri (con la possibilità di selezionarli tutti);
- l'edizione o le edizioni del testo (con la possibilità di selezionarli tutti);
- il secolo o i secoli (con la possibilità di selezionarli tutti);
- l'anno o gli anni (con la possibilità di selezionarli tutti).

Al termine di questa fase, all'utente comparirà a schermo una lista di tutti i valori (le entità) che è possibile inserire nella query di ricerca. Infatti, il fine ultimo del database è quello di permettere ricerche sia interne ad un singolo documento sia trasversali ai fini comparativi. Non dimentichiamoci che l'obiettivo finale della ricerca è studiare l'evoluzione diacronica di questi tipi di testi per poter poi formulare una definizione del genere testuale, in questo caso, delle GEI. Inoltre, l'intenzione è anche quella di proporre un modello di studio standardizzato da estendere anche ad altri testi analoghi ma, per esempio, pubblicati in epoche differenti o trattanti lingue diverse dall'italiano e lo spagnolo.

Volendo mostrare i risultati dell'interrogazione in forma tabellare nell'interfaccia utente, quest'ultimo potrà scegliere quali colonne visualizzare ma sempre a partire dalla forma o dal contenuto dei libri, come già anticipato poc'anzi. Infatti, sarà possibile visualizzare le entità della macrostruttura (quindi il layout), come, per esempio, la presenza o meno di capitoli; oppure entità della microstruttura (quindi il contenuto), come, per esempio, i tempi verbali usati o gli esempi inseriti; o una combinazione di colonne che preveda entrambi i criteri di studio (quindi sia il layout sia la forma). Per una data ricerca verrà data la possibilità di mostrare le colonne "capitoli", "esempi", "tempi verbali" e "parte del discorso trattata". Per ogni colonna si potrà applicare un ulteriore filtro: si possono escludere tutte le righe in cui il valore "esempi" è nullo (quindi mantenere solo le righe in cui vi sia un esempio valido) oppure le colonne in cui l'argomento trattato include "aggettivi".

Effettuato questo secondo processo di filtraggio, all'utente comparirà un primo risultato dell'interrogazione fin qui impostata. I risultati potranno essere mostrati sia attraverso l'elenco dei risultati (per esempio in forma tabellare, quindi le righe che risultano vere dai parametri di ricerca impostati), sia attraverso una forma di raggruppamento simile a una tabella pivot (quindi potendo espandere i risultati partendo dalle colonne che più raggruppano i risultati, ovvero quelle con le occorrenze simili più frequenti, come potrebbero esserlo la colonna "capitolo" o "esempio"), al momento ancora in fase di definizione. Per ogni "cella", si potranno scegliere le informazioni ulteriori in base alla colonna in cui si trovano: per esempio l'argomento trattato potrebbe essere una parte del discorso chiamata dall'autore con un nome desueto e l'utente potrebbe vedere tale nome desueto ed essere interessato al nome attuale, o viceversa. Queste informazioni variano da colonna a colonna e possono essere espanse (visualizzate) oppure aggiunte a ulteriori filtri (per esempio, limitare la ricerca all'argomento chiamato solo col nome desueto, oppure ampliare la ricerca dell'argomento dal nome desueto a quello più comune).

Sarà possibile fare un report della ricerca: per esempio il numero di occorrenze vere (righe risultanti, tra i testi/autori scelti dove si concentrano più risultati, qual è la percentuale di occorrenze vere per ognuno dei capitoli emersi dalla ricerca eccetera).

#### 4. RINGRAZIAMENTI

Il presente lavoro rientra nelle iniziative portate avanti dal Centro di Studi in Terminologia Computazionale (CENTRICO) dell'Università di Padova e nelle direzioni di ricerca dell'Infrastruttura Comune Italiana per le Risorse e le Tecnologie Linguistiche CLARIN-IT.

#### BIBLIOGRAFIA

- [1] Calero Vaquera, María Luisa. *Historia de la gramática española: 1847-1920* de A. Bello a R. Lenz. Biblioteca románica hispánica. 2, Estudios y ensayos. Madrid: Gredos, 1986.
- [2] Castillo Peña, Carmen. «Epigrama: Un portal para la edición digital de textos gramaticales», 2020.
- [3] Fornara, Simone. *La trasformazione della tradizione nelle prime grammatiche italiane (1440-1555)*. Roma: Aracne, 2013.
- [4] Gallina, Anna Maria. *Prime grammatiche spagnole ad uso degli Italiani (sec. 16.)*. Venezia: Tip. commerciale, 1975.
- [5] Gómez Asencio, José Jesús. *El castellano y su codificación gramatical. De 1492 (A. de Nebrija) a 1611 (John Sanford)*. Vol. 1. Beltenebros. Burgos: Fundacion Instituto Castellano y Leones de la Lengua, 2006.
- [6] *El castellano y su codificación gramatical. De 1700 a 1835*. Vol. 3. Beltenebros. Burgos: Fundacion Instituto Castellano y Leones de la Lengua, 2011.
- [7] *El castellano y su codificación gramatical. De (B. Jiménez Patón) 1614 a 1697 F. Sobrino*. Vol. 2. Beltenebros. Burgos: Fundacion Instituto Castellano y Leones de la Lengua, 2006.
- [8] Kukenheim, Louis. *Contributions a l'histoire de la grammaire italienne, espagnole et française a l'époque de la Renaissance*. Amsterdam: Noord-Hollandsche Utgevers-Maatschappij, 1932.
- [9] Ramajo Caño, Antonio. *Las gramáticas de la lengua castellana desde Nebrija a Correas*. Salamanca: Universidad de Salamanca, 1987.
- [10] San Vicente, Félix, e Hugo E. Lombardini. «Gramáticas de español para itálofonos de los siglos XVI-XVIII: elementos lexicográficos y componentes técnico-científicos». In *Lengua de la ciencia e historiografía*, a cura di José Ignacio Pérez Pascual e Cecilio Garriga Escribano. Anexos de Revista de lexicografía 35. A Coruña: Universidade da Coruña, 2016.
- [11] Trabalza, Ciro. *Storia della grammatica italiana*. Bologna: A. Forni, 1963.

# Un'ontologia per la classificazione semantica dei dizionari storici: l'esempio del LEI Digitale

Alessandro Alfredo Nannini<sup>1</sup>

<sup>1</sup> Università di Vienna, Austria – alessandro.alfredo.nannini@univie.ac.at

## ABSTRACT

Il poster intende presentare le possibilità di riutilizzo, nell'ambito della lessicografia storico-etimologica italo-romanza, di un sistema concettuale costruito a partire dalla classificazione semantica del lessico anglosassone. Infatti, il progetto di mappatura semantica del *Lessico Etimologico Italiano (LEI)* utilizza come base noemica la tassonomia dell'*Historical Thesaurus of English*, opportunamente convertita in ontologia (OWL) e ampliata in base alle esigenze di classificazione del lessico trattato nel *LEI*. I risultati ottenuti fanno ben sperare anche nella prospettiva di un utilizzo comune da parte dei diversi progetti di lessicografia storica.

## PAROLE CHIAVE

Historical Lexicography; Italian Lexicography; Ontologies; Semantics; Onomasiology

## 1. LESSICOGRAFIA ONOMASIOLOGICA NELL'ERA DIGITALE

La lessicografia onomasiologica classifica il materiale lessicale all'interno di sistemi concettuali che intendono rappresentare i concetti della realtà extralinguistica verbalizzati in una o più lingue. Tra gli esempi di questo approccio nei secoli, ricordiamo *La Fabrica del Mondo* di Francesco Alunno, il *Thesaurus of English Words and Phrases* di Peter Mark Roget e, in epoca più recente, l'*Historical Thesaurus of English* che classifica per concetti l'intero lessico anglosassone. Considerata spesso di peso minore rispetto alla "sorella" semasiologica, la lessicografia onomasiologica mostra un rinato vigore nell'era digitale.

Sin dai primi anni Ottanta del secolo scorso, la possibilità concreta di utilizzare strumenti informatici per la compilazione e la pubblicazione dei dati lessicografici ha rivoluzionato la concezione del dizionario che, da oggetto esclusivamente cartaceo contraddistinto da inevitabili vincoli di consultazione, è venuto a configurarsi come prodotto elettronico caratterizzato dalla multifunzionalità: «all the information which is stored in the dictionary is both technically and logically speaking available» (Zimmermann 1983, 279). Nel contesto multifunzionale si moltiplicano le possibilità di consultazione del materiale lessicografico e si neutralizzano le opposizioni classiche della lessicografia tradizionale, come, ad esempio, quella tra dizionario semasiologico e dizionario onomasiologico (cfr. Kipfer 1986; Calzolari 1988).

## 2. LESSICOGRAFIA ONOMASIOLOGICA DIGITALE IN AMBITO ROMANZO

La lessicografia romanza ha indugiato per molti anni sulle soglie del digitale. Soltanto a partire dal primo decennio del 2000 sono state concepite nuove opere native digitali, così come altri progetti in corso hanno intrapreso il faticoso processo di digitalizzazione; in entrambi i casi, la multifunzionalità digitale si è offerta come un'opportunità per garantire agli utenti diverse chiavi di lettura dei dizionari, compresa quella semantica (cfr. *DÉRom; Verba Alpina*). Nell'ultimo decennio la lessicografia storica romanza è entrata in contatto con le tecnologie del *Semantic Web*. Tittel/Chiarcos 2018 hanno illustrato il processo di conversione del *Dictionnaire Étymologique de l'Ancien Français* (DEAF) in Linguistic Linked Open Data (LLOD) modellati con *OntoLex-Lemon* in prospettiva di una mappatura semantica coi concetti presenti nelle ontologie (ad es. DBpedia). Anche il progetto *Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan* (DiTMAO) utilizza i LLOD per l'annotazione semantica di un dizionario, nel quale «l'organizzazione ontologica, naturalmente delimitata a determinati sistemi noemici è stata considerata la più adeguata a strutturare la complessa realtà terminologica: ad ogni concetto sono collegati i principali tipi lessicali» (Corradini/Mensching 2013, 117; cfr. anche Weingart/Giovannetti 2016). Comune ai diversi dizionari è la necessità di un sistema concettuale (sistema noemico) di riferimento, cioè di una tassonomia per la classificazione del lessico. Tale esigenza è tradizionalmente ben nota alla lessicografia onomasiologica: nel 1953 i romanisti Rudolf Hallig e Walter von Wartburg hanno pubblicato il *Begriffssystem als Grundlage für die Lexicografie*, un sistema di concetti della realtà extralinguistica (rappresentati in lingua francese) per la strutturazione dei dizionari onomasiologici. Tuttavia, il *Begriffssystem* presenta delle difficoltà di applicazione che lo hanno portato a essere scartato dai progetti di classificazione semantica più recenti (cfr. Tittel et al. 2020).



La costruzione di un sistema concettuale nel quale classificare una grande quantità di dati linguistici afferenti ai diversi assi di variazione, come quelli registrati dalla lessicografia storica, è una delle grandi sfide sul banco di molti progetti in corso d’opera.

### 3. UN’ONTOLOGIA PER L’ANNOTAZIONE SEMANTICA DEL *LEI DIGITALE*

L’annotazione semantica mediante l’utilizzo di un sistema ontologico è uno degli obiettivi portati avanti nel processo di digitalizzazione del *Lessico Etimologico Italiano* (LEI), il più ricco dizionario storico-etimologico delle varietà italo-romanze (cfr. Prifti 2022, 302-303). Per il conseguimento di questo fine, il *LEI Digitale* ha optato per il riutilizzo della tassonomia dell’*Historical Thesaurus of English* (HTE), rovesciamento onomasiologico dell’*Oxford English Dictionary* (OED). L’HTE classifica il lessico inglese su dodici livelli (sette di categorie principali e cinque di sottocategorie) «in a taxonomy which begins with the most general ways of expressing a concept and moves hierarchically downwards to the most specific» (Alexander/Kay 2023).

La mappatura semantica del LEI, operata su tre livelli strutturali del dizionario (etimo/lemma dell’articolo, sezioni semantiche interne all’articolo, forme italo-romanze in articolo), mette in evidenza le possibilità di riutilizzo di un sistema concettuale costruito a partire dalla classificazione del lessico di un’altra lingua (l’inglese). Tutte le categorie del sistema sono state modellate come classi ontologiche in OWL (Figura 1) e, quando la classificazione semantica del lessico italo-romanzo lo ha richiesto, si sono operate delle aggiunte alla tassonomia inglese che si configura così come una solida base (*Grundlage*, per richiamare Hallig e Wartburg) ampliabile e non come una griglia fissa da imporre al lessico.

Lessemi e classi ontologiche si interconnettono in un grafo che può essere indagato sia semasiologicamente che onomasiologicamente. Questa modellazione si presta bene sia alla rappresentazione della multidimensionalità dei significati linguistici che alla visualizzazione di quei *chunks of knowledge* (‘pezzi di conoscenza’) cognitiva ed enciclopedica che si relazionano con il lessico. Inoltre, grazie al formato digitale si superano anche le rigide strutture dei dizionari onomasiologici classici che, pur classificando semanticamente il lessico, non sono in grado di rappresentar efficacemente la complessità delle relazioni che intercorrono tra i suoi elementi.

La soluzione proposta per il LEI è da considerarsi anche nell’ottica di interconnessione semantica tra le diverse opere di lessicografia storica (cfr. Giuliani/Molina Sangüesa 2020; Prifti 2022; Nannini in stampa).

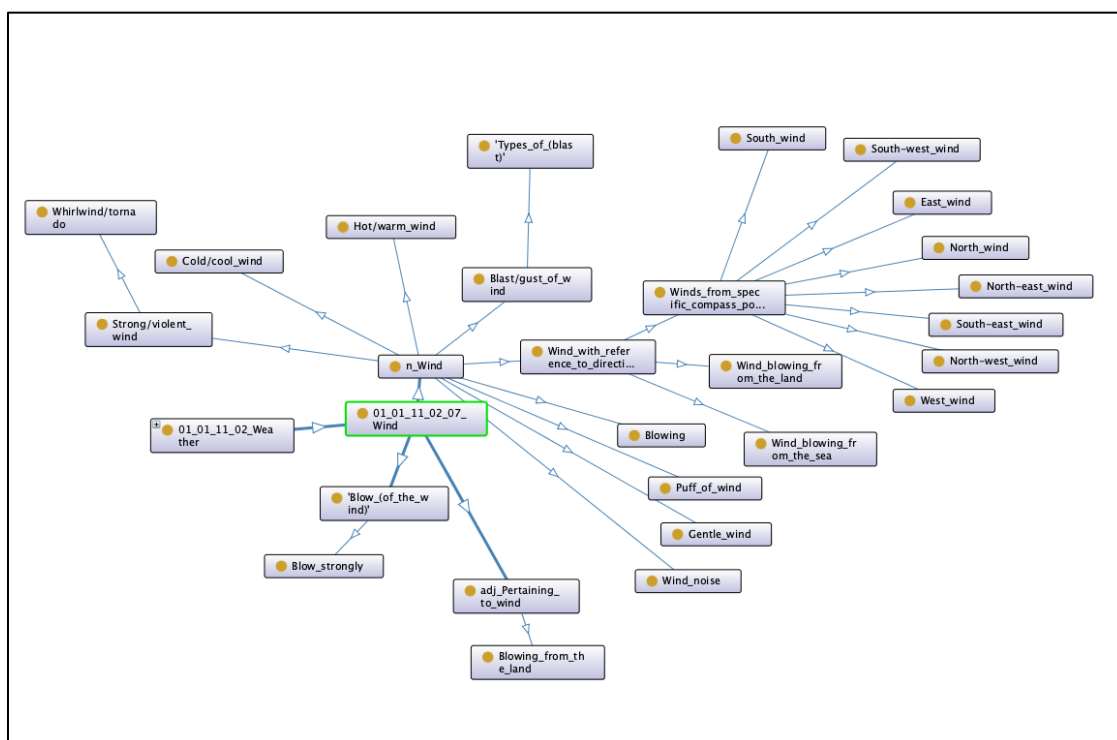


Figura 28. Alcune categorie relative al vento visualizzate con OntoGraf (Protégé).

### 4. CENNI SULLA MAPPATURA SEMANTICA DEL *LEI DIGITALE*

Le informazioni lessicali del *LEI Digitale* sulle quali si concentra il nostro lavoro sono attualmente codificate in XML. Ogni elemento del dizionario che contiene informazioni semantiche viene collegato al concetto (o ai concetti) di riferimento

mediante l'utilizzo dell'attributo *@corresp* con l'URL della classe ontologica come valore d'attributo (cfr. Bowers/Declercq 2016).<sup>1</sup>

Dopo una fase di ricerca volta a individuare il sistema concettuale più adatto per il LEI e la successiva fase di conversione del sistema in OWL, nel 2022 ha preso avvio la fase di mappatura del significato degli etimi presenti nel materiale digitalizzato, relativamente alla lettera A e B del dizionario. Contemporaneamente si è svolta la mappatura integrale degli articoli afferenti alla famiglia lessicale composta da *boreālis*, *boreās/borrāsboreus* e *borinus* (Figura 2).

Nella prossima fase (seconda metà del 2023) il materiale mappato verrà messo a disposizione, insieme al sistema concettuale, sul sito del *LEI Digitale*.<sup>2</sup>

<p><b>boreālis</b></p> <p>'boreale, settentrionale'</p> <p>Concetto Etimo: <a href="#">adj_North</a></p> <p>II.1.</p> <p>Concetto: <a href="#">adj_North</a></p> <p>It. <i>boreale</i> agg. 'settentrionale, nordico' (dal 1304, Plutarco Volg, B; TB; Crusca 1866; Zing 1998)</p> <p>venez. <i>boreal</i> (Saverien 1769)</p> <p>nap. <i>boreale</i> (1689, Fasano, Rocco)</p> <p>Concetto: <a href="#">n_Aurora_borealis</a></p> <p>it. <i>aurora boreale</i> 'meteora luminosa delle altissime regioni dell'atmosfera' (dal 1623, Galilei, B; TB; Crusca 1866; Zing 1998)</p> <p>Concetto: <a href="#">adj_Causing_wonder_astonishing</a></p> <p><i>boreale</i> agg. 'fantastico, eccezionale' ("ling.giovanile" Forconi 1988)</p>
---

Figura 29. Una parte dell'articolo *boreālis* (LEI, VI, 1070) ordinato per concetti.

Il poster che proponiamo in questa sede illustra il workflow seguito nelle diverse fasi del lavoro, presenta alcuni risultati parziali e offre una panoramica sulle possibilità di indagine linguistica offerte dal nuovo strumento.

## BIBLIOGRAFIA

- [1] Alexander, Marc, e Christian Kay. «Classification». In *About the Historical Thesaurus of English*. Glasgow: University of Glasgow, 2023. <https://ht.ac.uk/classification/>
- [2] Alunno, Francesco. *La Fabrica Del Mondo*. Venezia, 1548.
- [3] Bowers, Jack, e Thierry Declercq. «TEI and LEMON: a comparative study on the lexical encoding and interoperability». *TEI conference and members' meeting 2016*, Settembre 2016, Vienna, Austria, <https://inria.hal.science/hal-02921955>
- [4] Buchi, Eva, e Wolfgang Schweickard, a c. di. *Dictionnaire Étymologique Roman (DÉRom)*. Nancy: ATILF, 2008-. <http://www.atilf.fr/DERom>
- [5] Calzolari, Nicoletta. «The dictionary and the thesaurus can be combined». In *Relational models of the lexicon. Representing knowledge in semantic networks*, a cura di Martha Walton Evens, 75–96. New York: Cambridge University Press, 1988.
- [6] Corradini, Maria Sofia, e Guido Mensching. «Nuovi aspetti relativi al “Dictionnaire de Termes Médico-botaniques de l’Ancien Occitan” (DITMAO): creazione di una base di dati integrata con organizzazione onomasiologica». In *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas*. 6-11 septiembre 2010, Valencia, a cura di Emili Casanova e Cesáreo Calvo, VIII:113–24. De Gruyter, 2013. <https://doi.org/10.1515/9783110300031.113>
- [7] Giuliani, Mariafrancesca, e Itziar Molina Sangüesa. «Hacia una taxonomía integrada en la redacción y revisión de diccionarios históricos». *Bollettino dell’Opera del Vocabolario Italiano*, fasc. XXV (s.d.): 325–74.
- [8] Kay, Christian, Marc Alexander, Fraser Dallachy, Jane Roberts, Michael Samuels, e Irené Wotherspoon, a c. di. *The Historical Thesaurus of English*. 2a ed., version 5.0. University of Glasgow, 2023.
- [9] Kipfer, Barbara Ann. «Investigating an onomasiological approach to dictionary material». *Dictionaries: Journal of the Dictionary Society of North America*, fasc. 8 (1986): 55–64.

<sup>1</sup> La serializzazione del materiale lessicografico in RDF e l'utilizzo di *OntoLex-Lemon* sono state prese in considerazione per una fase successiva del progetto.

<sup>2</sup> <https://online.lei-digitale.it>

- [10] Krefeld, Thomas, e Stephan Lücke, a c. di. *Verba Alpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit*. München, online, 2014. <https://dx.doi.org/10.5282/verba-alpina>
- [11] Nannini, Alessandro Alfredo. «La mappatura semantica del Lessico Etimologico Italiano (LEI). Possibilità, metodi e prospettive» In *Digitale romanistische Sprachwissenschaft. Akten des Romanistischen KolloquiumsXXXIV*, a cura di Lidia Becker et al. Tübingen, Narr, in corso di stampa.
- [12] Prifti, Elton. «Il LEI digitale. Un resoconto, con particolare attenzione alla dialettologia». In *Lessicografia storica dialettale e regionale*, a cura di Michele Cortelazzo, Silvia Morgana, e Massimo Prada, 293–314. Firenze: Franco Cesati, 2022.
- [13] Prifti, Elton, e Wolfgang Schweickard, a c. di. *LEI. Lessico Etimologico Italiano*. Fondato da Max Pfister. Wiesbaden: Reichert, 1979-. *LEI Digitale*: <https://lei-digitale.it>
- [14] Roget, Peter Mark. *Thesaurus of English Words and Phrases: Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition*. Cambridge Library Collection - Linguistics. Cambridge: Cambridge University Press, 2014. <https://doi.org/10.1017/CBO9781107448735>
- [15] Tittel, Sabine, e Christian Chiarcos. «Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the Dictionnaire étymologique de l'ancien français with OntoLex-Lemon». In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, a cura di Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, et al., 58–66. Miyazaki, Japan: European Languages Resources Association (ELRA), 2018.
- [16] 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. Marseille, France: European Language Resources Association, 2020. <https://aclanthology.org/2020.ldl-1.1>
- [17] von Wartburg, Walther., e Rudolph Hallig. *Begriffssystem als Grundlage für die Lexikographie, Versuch eines Ordnungsschemas*. 2a ed. *Abhandlungen der deutschen Akademie der Wissenschaften zu Berlin*. Berlin: Akademie-Verlag, 1963.
- [18] Weingart, Anja, e Emiliano Giovannetti. «A Lexicon for Old Occitan Medico-Botanical Terminology in Lemon». In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016)*, Heraklion, Greece, May 30th, 2016, a cura di Isabelle Draelants, Catherine Faron-Zucker, Alexandre Monnin, e Arnaud Zucker, 1595:25–36. *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. <https://ceur-ws.org/Vol-1595/paper3.pdf>
- [19] Zimmerman, Herald H. «Multifunctional Dictionary». In *The possibilities and limits of the computer in producing and publishing dictionaries: proceedings of the European Science Foundation Workshop*, Pisa, 1981, a cura di Antonio Zampolli e Amadeo Cappelli, 279–88. Pisa: Giardini, 1983.

# How many oral archives are in your home?

## Piloting a new Tuscan census in the Gra.fo Reloaded project

Duccio Piccardi<sup>1</sup>, Silvia Calamai<sup>2</sup>

<sup>1</sup> University of Siena, Italy – duccio.piccardi@unisi.it

<sup>2</sup> University of Siena, Italy – silvia.calamai@unisi.it

### ABSTRACT

In this paper, we describe the pilot phase of a new census of Tuscan oral archives promoted by the Gra.fo Reloaded project. Building on a long research tradition, the Gra.fo Reloaded census strives to take a step forward by a) updating the contents of previous similar endeavours; b) promoting contemporary initiatives to strengthen the oral archive research community; c) adapting well-established pre-testing procedures of National census agencies to its specific context. 13 individuals of diverse disciplinary backgrounds were asked to fill in the census form while thinking aloud; then, they completed an Italian reduced version of the Questionnaire of User Interface Satisfaction (QUIS). The results make a solid argument for running similar procedures in census construction processes and being aware of the features of our survey tools.

### KEYWORDS

Gra.fo Reloaded; Oral Archives; Census; Cognitive interview; QUIS

## 1. INTRODUCTION

Inaugurating the first electronic issue of *Oral History Forum d'histoire orale*, Canadian archivist Wilma MacDonald provokingly asked her readers “what happens to the oral history you create?”, while encouraging them to store their oral documents in suitable repositories for the sake of accessibility and long-term preservation [1]. Far from being a mere outburst of archival idealism, MacDonald’s call stemmed from her first-hand experience as an oral-archive survey conductor: comparing data from two National censuses launched in 1984 and 1994, she witnessed a steady increase in the production of tape-recorded documents [2]. This situation, in turn, urged to make both institutions and researchers aware of the importance of their materials, in order to reduce the risk of irreparable losses. Indeed, the digital revolution makes the increase rate of oral document production of the past decades appear as small talk. In this new milieu, the advantages of digital orality cannot be fully achieved if the broadened availability of tools for high-quality recording outpaces our capability of knowing, archiving, and safeguarding oral documents while making them easily available ([3] and, mutatis mutandis,[4]).

As for any other form of intangible cultural heritage, a first step on the pathway to the preservation of orality is acquiring knowledge of the existing materials, and, therefore, conducting and updating surveys, which is instrumental to the planning of long-term actions(e.g.,[5]). However, this is not perse a small feat, since archive surveys are time-expensive activities for both the conductor and the respondents, who are not always in the right conditions to answer with quality information, if not to answer at all ([2]; [6]) – not to mention that the current digital “data flood” [7] most certainly implies tightened update intervals. A possible solution to this conundrum is the projecting of censuses which are not just periodic representations of the status quo, but systematically linked to permanent structures conceived for the spontaneous participation of the target population in any time and space. This is, for example, the approach advanced by a work-in-progress by the Italian Istituto Centrale per i Beni Sonori e Audiovisivi [8], which will host the census form on a long-lasting website designed for regular participative updating. In the same spirit – albeit more local in scope, another oral archive survey is currently underway: the Tuscan Gra.fo Reloaded census. After a brief introduction to the general project structure (§2), this paper will focus on the census planning, from the selection of the sources and the loci of constitution of a stable community of Tuscan oral archivists (§3), the formal guidelines (§4), up to the results of a pre-test phase (§5). Lastly, over all conclusions will be drawn(§6).

## 2. GRA.FO RELOADED

Gra.fo Reloaded is a project financed by Regione Toscana (2022-2024) and coordinated by Siena University in collaboration with Fondazione Sistema Toscana, Soprintendenza Archivistica e Bibliografica della Toscana, Ecomuseo della Montagna Pistoiese, Istituto di Linguistica Computazionale “Antonio Zampolli” (Pisa National Research Council), and CLARIN-IT. The project encompasses three main research lines. Firstly, as the name implies, Gra.fo Reloaded aims to establish a pipeline for the revitalization of a previous Tuscan project on oral archives, Grammo-foni. Le soffitte dellavoce (Gra.fo; Scuola Normale Superiore of Pisa and University of Siena, 2007-2013 [9]). During the course of its activities, Gra.fo surveyed, digitized, described, and made available to the public around 30 oral archives (3000 hours of recordings). However, its web access portal is no longer maintained and currently closed. For this reason, Gra.fo Reloaded plans to study the feasibility of transferring the original data on a much newer (and hopefully durable) infrastructure, Archivio Vi.Vo. ([10]; see below, §3). In this process, Gra.fo Reloaded also relies on a previously uninvolved partner (Fondazione Sistema Toscana) in order to further amplify the reach of its activities and improve their sustainability over time. The second project objective consists in the new Tuscan census, which will be discussed in detail below. Thirdly, Gra.fo Reloaded focuses on the re-use of oral archival materials for the sake of participative territory promotion. The Gra.fo documents recorded on the Apennine range in the Province of Pistoia will be analysed and adapted to an augmented-reality geolocated audio tour (e.g., [11]) revolving around the areas covered by the local ecomuseum, thus realizing a long-planned extension of the original enterprise [12]. Citizens of the Pistoiese range will be consulted for both the itinerary sketching process and the gathering and production of additional oral documents. These three seemingly disconnected activities are, in actual fact, deeply intertwined. For example, the feasibility inquiry for the Gra.fo – Archivio Vi.Vo. transfer is being conducted using materials from the Pistoiese range as a case-study, thus also serving the third objective of the project. Moreover, the census is connected to the Archivio Vi.Vo. platform, alongside other contemporary initiatives on oral archives in Italy, as it will be shown in the next paragraph.

## 3. A TUSCAN COOKBOOK FOR AN ORAL ARCHIVE CENSUS: THE INGREDIENTS

As [10] recently stressed, the search tradition on oral archives in Tuscany is rather sound, counting the works of Giovanni Contini at the Soprintendenza Archivistica e Bibliografica della Toscana (who later co-authored [13]), seminal publications on specific topics [14] and, of course, dedicated censuses, among which [15] was the first uncovering 124 Regional oral archives between late 1999 and early 2004. The aforementioned original Gra.fo survey was a partial update and extension of [15]: indeed, this first Tuscan census focused on oral history and anthropology, while leaving aside archives pertinent to linguistics subdisciplines. To these aims, addition [15], Gra.fo surveyed other Italian oral archive censuses (namely [16,17]) and then added to the mix of its form design other linguistics-related elements. In order to take a step further, the Gra.fo Reloaded census construction procedure repeated this pipeline, also adapting elements from [18] and [19], the latter of which, despite not being a census *sensu stricto*, helped formulating sections related to digital documentation (e.g., file formats), which was mostly left out previous similar efforts. Moreover, references to technology and the internet were also added from scratch: for example, an extensive archive content topic list was directly imported from Gra.fo while adding more contemporary keywords (e.g., “videogames”, “social networks”). This ensured not only an overall better topic coverage, but also some level of compliance with emerging speech research protocols (such as gamified ones; e.g. [20]). We also strived to achieve a good balance between standardization and accessibility of the deployed terminology. For example, we substituted the four-level carrier physical condition gradings scale from [18] with the one from [21] (alongside an adaptation of the respective semantic tags), which, being dedicated to collectors, might be more acknowledged by the general public. As we briefly mentioned above (§1), the Gra.fo Reloaded census makes explicit reference to other contemporary Italian projects on oral archives for the sake of community building. Firstly, the census describes to the respondent the software developed through the ArchivioVi.Vo. project (Regione Toscana 2019-2021; University of Siena, Soprintendenza Archivistica e Bibliografica della Toscana, Istituto di Linguistica Computazionale “Antonio Zampolli” - Pisa National Research Council - and CLARIN-IT, Ecomuseo del Casentino): an easy-to-use online platform with an integrated workflow guiding the archivist from the creation of a long-term preservation digital copy to the finalization of the individual sound documents contained by the carrier through interfaces dedicated to audio restoration and content description – all being compliant to international standards [10]. Secondly, the respondent is informed of [22], a *Vademecum* on oral document creation, description, preservation, and valorisation jointly written by a heterogeneous group composed by members of the CLARIN-IT consortium, National institutions, and scientific associations. With respect to both these tools, the respondent can declare his/her interest in dedicated sections and be kept up-to-date of related initiatives via future mailing lists. Through an ideal consistent reach of [8] and our census, the Tuscan research community will be soon acquainted with several bootstrap tools related to the investigation of oral sources (an

accessible text for learning fundamental concepts on disparate aspects of the workflow; a participative perpetual census of the gathered materials; an open-access, intuitive software for oral material processing and storage), potentially triggering a snowball effect of lasting awareness (see also the “crowdsourcing” idea of the original Gra.fo project [23]). Of course, the close contact with Archivio Vi.Vo. and the Vademecum left other structural marks on the Gra.fo Reloaded census. While a peculiar attention paid to restoration, digitization, legal issues, and related archival documents stem from the former, the latter, through the consequently constituted Tavolo permanente per le fonti orali, allowed us to fruitfully discuss with [8] during the respective development phases.

To sum up, this first phase of source selection, revision, and integration left us with the following census macro-sections: physical carriers; digital documents; other archive features (archival order, description, long-term preservation); document contents; other related documents; legal issues; publications; previous knowledge of pertinent topics; declarations of interest; personal data (of all the involved subjects) 2. Now that all these census ingredients were placed on the worktable, we had to look for a viable strategy for combining them and assessing their goodness. To this aim, we searched for foreign recipes.

#### 4. THE RECIPE

Literature on survey design is abundant. Here we will briefly cover two main areas of our interest, namely the modes of data collection and the rationale behind question structuring. In the mid Nineties, survey scientists began to explore the potential of the world wide web and its sampling problems [24]: given the limited diffusion of the internet, the subset of the population reachable through it presented biased features. The late adoption of web-based procedures in Italian oral archive surveys goes well beyond the sake of statistical finesse. Among the sources mentioned in §3, [18] was the first to evaluate a multimode data collection strategy involving the web, the complete adoption of emails being hindered by their limited diffusion among local institutions. When compared to European standards, Italy has a history of subpar internet access, especially in famiglie consisting of 65+ years old people and residing in places far from major urban areas [25;26]. An Italian oral archive census cannot forgo these two demographic subsets: recall that the biggest archive discovered in Tuscany by [15] is possessed by a religious community in the country side of Grosseto. The decision of previous endeavours to rely on phone calls and paperworks was not a matter of achieving a representative sample: it was a matter of achieving a successful census. Despite this peculiar context, oral archive survey conductors fully transitioned to web-based approaches in more recent times [19; 8]. [19] relied on Google Forms, an online survey tool which caught the eye of the scientific community since its first versions because of its free availability, intuitiveness, and functionalities –in particular, its skip logic and data plotting features [27]. More recently, [28] also highlighted the low cognitive workload required to use the tool and the familiarity that a good number of internet users has with it –these properties being probably interrelated from the viewpoint of text legibility [29]. These qualities convinced us to plan a sequential multimode strategy consisting of a first, time non-expensive, campaign distributing a Google Forms implementation of the census via email followed by an ad hominem follow-up (via phone or in person) to reduce non-response rate. In order to keep an eye on potential mode effects on responses, the conductor will keep track of the circumstances of questionnaire completion [30]. With respect to the other topic of interest, previous research highlight that question design choices are hardly universally right or wrong, suggesting that knowledge of the specific trade-offs is essential in tailoring an effective survey. A good exemplification of this point is the line of inquiry on check-all-that-apply (CATA) versus forced-choice (FC) question format. It is generally assumed that the latter leads to more attentive answers and Higher endorsement rates of the individual options [31]. However, the sheer advantages of FCs are probably too small in comparison to the higher CATA perceived usability [32]. Moreover, as our census form is rather extensive, time completion times are also a concern [33], leading us to prefer CATAs for questions with long lists of options (such as the content topic selection commented above), while eventually trying to gather more precise information in follow-ups. Even though an extensive discussion on all the design features of the census is out of the scope of this paper, we will list here a series of elements with pertinent bibliographical references. We tried to write our questions with a concise style [34], but with frequent parenthetical clarifications (definitions, formatting requirements, etc.) positioned after the question [35]; in compliance with the implementation of Google Forms skip logic, the number of questions per census page was kept moderated [36], and a progress indicator was provided for transparency (but see [37]). Lastly, given the absence of consensus on the issue, we followed the general tendency of placing demographic questions on the involved subjects after the other sections ([38] for a recent short review). At the end of this assembling phase, we ended up with a preliminary Google form composed of 47 content pages and 98 total questions (mean 2.08 questions per page). This total estimate might be seen as intimidating; however, the implementation of Forms skip logic progressively reduces the completion burden alongside with the decrease

of surveyed archive complexity, leaving the minimal cases with approximately half of the total questions to fill in. In any case, before the launch of the survey, we needed to answer a fundamental question: how does this new census recipe taste?

## 5. USABILITY TESTING

In order to take a first look at how potential users would interact with our new census, we run a usability test inspired by the guidelines of two National Agencies with a tradition of formalizing and expounding their procedures to the scientific community: Statistics New Zealand (e.g.,[39]) and, in particular, the United States Census Bureau (e.g.,[40]). A two-step protocol was envisioned, consisting of a qualitative cognitive interview (for accuracy) followed by a quantitative questionnaire on user satisfaction.

*Participants and settings.* While a final quota of participants was not decided a priori, we opted to stop testing after 13 individual runs, since the number of new information and suggestions emerging from the procedure had evidently become stagnant. The recruited participants were acquaintances of the first author, did not have any prior archival knowledge, and had very diverse education levels (from high school diploma to university professor) and involvement with oral documents (music researchers and enthusiasts, linguists, psychologists, etc.). The participants were interviewed in private homes or online via video calls.

*Materials.* The preliminary version of the census extensively discussed in §§3-4 was the object of the inquiry. In order to have a quantitative assessment of user satisfaction, we adopted a multi-purpose tool, the Questionnaire of User Interface Satisfaction (QUIS) [41], in a 11 (specific) + 3 (overall reactions) items version reduced by the United States Census Bureau [42]. The reduced QUIS was translated into Italian (through content adaptation) and implemented in another Google form, with a 5-point Likert scale (with various semantic tags, following the general rationale of higher scores equal desirable properties) associated to each one of the questions.

*Procedure and analysis.* The experimenter sent the link with the Google census form to the participant and, if the test was conducted online, asked him/her to share the screen. Participants were requested to fill in the census as any other user of the future campaign while thinking aloud. In order to accustom them to this practice, we asked them to answer the training question “how many windows are in your home?” [43]; then, once the participant had warmed up, the experimenter took note of their interactions with the census form and their pertinent comments while trying to keep his active verbal interventions limited to explicit requests. In case participants declared that they did not possess any oral documents, the test was run using the following vignette: “imagine that you have a very unique collection of recordings of your favourite musician’s live performances and rehearsals”. After the completion of the form, the experimenter sent a second Google link with the QUIS to the participant asking him/her to provide an answer to each question. Each individual session lasted from around 30 mins. To 1 hour. Given the very limited number of responses, the analysis of the QUIS was conducted using the integrated descriptive statistics tool of Google Forms.

*Results: Cognitive interviews.* 44 distinct points of discussion emerged from the 13 interviews, with a mean of 6/7 comments per interview (and, thus, a noticeable overlap). A very common observation pertained to an off-putting formatting derived from a hasty question copy-paste from a development textual document. The other points can be roughly divided in two macro-categories: 31 requests for clarification features (mostly about archiving terminology, but also a small number of procedural opacities) and 12 suggestions for question/option revisions (mostly missing options – potential document contents, specific subdisciplines, etc. ).

*Results: QUIS.* The means of the responses to the individual questions were overall acceptable: intuitiveness (4.3), organization of the information (4.1), terminological coherence (4), information arrangement on screen (4.3), adequate presence of sectoral terminology (4.2), character readability (5), ease of learning to operate the census (4.8), appropriateness to inexperienced users (3.9), perceived census completion speed (4.3), operation predictability (4.8), ease of performing response operations (4.7); overall impression (terrible-wonderful, 4; frustrating-satisfying, 4; difficult-easy, 4.6). An open-ended final suggestion box was rarely filled in and will not be commented here. Note that the only mean score below 4 pertains to the needs of non-specialists, which confirms the importance of the many requests for clarification retrieved during the cognitive interviews.

*Census re-building.* The two authors jointly evaluated the contents of the 44 points of discussions and critically examined the best integration strategy for each of them. Then, for the sake of traceability, the test census Google form was copied in another URL and consequently modified. The final census has 49 pages (preceded by a consent form) and 100 questions (mean 2.04 questions per page). As the reader might notice, we strived to implement content revisions in the pre-existing structure without adding too many new elements: in fact, the QUIS responses indicated that the original census length was deemed acceptable. The final census is available at <https://forms.gle/kwRJeacsdP3eXCo8>.



## 6. SUMMARY AND CURRENT WORK PROGRESS

Given the dramatic increase in data production, conducting oral archive surveys risks becoming a researcher's Sisyphus boulder. In this paper, we highlighted that a contemporary Italian census trend tries to mitigate this condition by connecting users to resources apt for self-learning, awareness developing, and autonomous intervention. In an attempt to make the knowledge gained through years of Tuscan oral archive tradition percolate into a new research community, the Gra.fo Reloaded census does its part. This new form is not only dedicated to non-archivists, but also structured in cooperation with them. Usability testing underlined the importance of the KISS (Keep It Simple, Stupid) principle in survey design, which is indeed not only limited to mere question length [34]. Even though technology offers easy solutions (such as skip logic features) to reduce the disruptive effects of burdensome numbers of survey items, terminological and procedural clarity must be a primary concern for conductors. In this regard, our pre-test phase highlights the advantages of relying on an external perspective to efficiently spot residual issues. At the present time, the census has been sent to over 350 addresses, including research associations, Municipality offices, ecomuseums, Ph.D. program coordinators, and individual researchers. Related information was also disseminated through newsletters and during events devoted to oral sources. Gra.fo Reloaded and its census caught the eye of the press and was covered by the daily newspaper *La Repubblica* [44]. In the meantime, we are beginning to narrow down our focus in order to plan the subsequent 'ad hominem' strategy of data collection.

## REFERENCES

- [1] MacDonald, Wilma. «What Happens to the Oral History You Create?». *Oral History Forum d'histoire orale* 27-28 (2007-2008):1-4.
- [2] MacDonald, Wilma. «Archival Theory and Oral History Documents». Master's Thesis, University of British Columbia, 1995.
- [3] Boyd, Doug. «Achieving the Promise of Oral History in a Digital Age». In *The Oxford Handbook of Oral History*, edited by Donald Ritchie, 285-302. New York: Oxford University Press, 2011.
- [4] Webster, Michael S., and Gregory F. Budney. «Sound Archives and Media Specimens in the 21st Century». In *Comparative Bioacoustics: An Overview*, edited by Charles Brown and Tobias Riede, 462-85. Oak Park: Bentham Science Publishers, 2017.
- [5] Yang, Hong. «The Pros and Cons about the Digital Recording of Intangible Cultural Heritage and Some Strategies». *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 40, no. 5/W7 (2015): 461-64.
- [6] Ferriter, Diarmaid. «Oral Archives in Ireland: A Preliminary Report». *Irish Economic and Social History* 25, no.1 (1998):91-5.
- [7] Jurgens, Charles. «Threats of the Data-Flood: An Accountability Perspective in the Era of Ubiquitous Computing». In *Archives in Liquid Times*, edited by Frans Smit, Arnoud Glaudemans, and Rienk Jonker, 196-210. The Hague: Stichting Archieffpublicaties, 2017.
- [8] Michetti, Giovanni. «Il censimento degli archivi sonori in Italia». Paper presented at *Un anno di 'Vademecum per il trattamento delle fonti orali'*, Rome, October 27, 2022.
- [9] Calamai, Silvia. «Ordinare archivi sonori: il progetto Gra.fo». *Rivista italiana di dialettologia* 35 (2011): 135-64.
- [10] Calamai, Silvia, Duccio Piccardi, Niccolò Pretto, Giovanni Candeo, Maria Francesca Stamuli, and Monica Monachini. «Not Just Paper: Enhancement of Archive Cultural Heritage». In *CLARIN. The Infrastructure for Language Resource*, edited by Darja Fišer, and Andreas Witt, 647-65. Berlin: De Gruyter, 2022.
- [11] Knight-Hill, Andrew, Brona Martin, and Aki Pasoulas. *Whispers of the Past*. 2021. <http://gala.gre.ac.uk/id/eprint/35928>
- [12] Pozzebon, Alessandro, Francesca Biliotti, and Silvia Calamai. «Places Speaking with Their Own Voices. A Case Study from the Gra.fo Archives». In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. EuroMed 2016*, edited by Marinos Ioannides, Eleanor Fink, Antonia Moropoulou, Monika Hagedorn-Saupe, Antonella Fresa, Gunnar Liestøl, Vlatka Rajčić, and Pierre Grussenmeyer, 232-39. Cham: Springer, 2016.
- [13] Contini, Giovanni, and Alfredo Martini. *Verba manent. L'uso delle fonti orali per la storia contemporanea*. Rome: La Nuova Italia Scientifica, 2013.
- [14] Di Piazza, Valeria, and Dina Mugnaini. *Io so' nata a Santa Lucia. Il racconto autobiografico di una donna toscana tra mondo contadino e società d'oggi*. Castelfiorentino: Società Storica della Valdelsa, 1988.
- [15] Andreini, Alessandro, and Pietro Clemente, eds. *I custodi delle voci. Archivi orali in Toscana: primo censimento*. Florence: Regione Toscana, 2007.
- [16] Barrera, Giulia, Alfredo Martini, and Antonella Mulè, eds. *Fonti orali. Censimento degli istituti di conservazione*. Rome: Ministero per i Beni Culturali e Ambientali, Ufficio Centrale per i Beni Archivistici, 1993.
- [17] Benedetti, Amedeo. *Gli archivi sonori: fonoteche, nastroteche e biblioteche musicali in Italia*. Genoa: Erga, 2002.
- [18] Cappelli, Fabrizio, and Antonella Rioda. «Archivi sonori in Toscana: un'indagine». *Musica/Tecnologia* 3 (2009): 9-69.
- [19] Galatà, Vincenzo, and Silvia Calamai. «Looking for Hidden Speech Archives in Italian Institutions». In *Selected Papers from the CLARIN Annual Conference 2018*, edited by Inguna Skadiņa, and Maria Eskevich, 46-55. Linköping: Linköping University Electronic Press, 2019.

- [20] Piccardi, Duccio, and Fabio Ardolino. «Gaming Variables in Linguistic Research. Italian Scale Validation and a Minecraft PilotStudy». In *Speaker Individuality in Phonetic sand Speech Sciences. Studi AISV8*, edited by Camilla Bernardasci, Dalila Dipino, Davide Garassino, Stefano Negrinelli, Elisa Pellegrino, and Stephan Schmid, 299–324. Milan: Officinaventuno, 2021.
- [21] Popoff, Martin. *Goldmine Record Album PriceGuide.6thEdition*.Iola: Krause,2009.
- [22] Coordinamento per le fonti orali Vademecum per il trattamento delle fonti orali. Rome, 2021. [https://www.aisv.it/vademecum\\_fonti\\_orali/Vademecum%2002.11.21.pdf](https://www.aisv.it/vademecum_fonti_orali/Vademecum%2002.11.21.pdf)
- [23] Biliotti, Francesca, and Silvia Calamai. «Web Dissemination of Tuscan Oral Archives». *Bulletin del'AFAS. Sonorités* 45 (2019). <https://doi.org/10.4000/afas.3439>
- [24] Baker, Reg. «Technology». In *A Meeting Place and More...: A History of the American Association for Public Opinion Research*, edited by Tom W. Smith, 117–44. Alexandria: American Association for Public Opinion Research, 2020.
- [25] Istituto Nazionale di Statistica. *Rapporto SDGs 2022. Informazioni statistiche per l'Agenda 2030 in Italia. 2022*. <https://www.istat.it/storage/rapporti-tematici/sdgs/2022/Rapporto-SDGs-2022.pdf>
- [26] Istituto Nazionale di Statistica. *Cittadini e ICT. Anno 2019. 2019*. <https://www.istat.it/it/files/2019/12/Cittadini-e-ICT-2019.pdf>
- [27] Travis, Lisa. «One of Many Free Survey Tools: Google Docs». *Journal of Electronic Resources in Medical Libraries* 7, no. 2 (2010): 105–14.
- [28]Dunn Lopez, Karen, Sheila M.Gephart, and Patricia E. Hershberger. «Using Online Survey Software to Enhance Rigorand Efficiency of Knowledge Synthesis Reviews». *WesternJournal ofNursingResearch*42, no.10 (2020): 838–45.
- [29]Zineddin, Abdulilah Z., Philip M.Garvey, Richard A. Carlson, and Martin T. Pietrucha. «Effects of Practice on FontLegibility».
- [30] *Proceedings of the Human FactorsandErgonomicsSocietyAnnualMeeting* 47,no.13 (2003):1717–20.
- [31] De Leeuw, EdithD.«To Mixor Not to Mix Data Collection ModesinSurveys». *Journal of Official Statistics* 21, no.2 (2005): 233–55.
- [32] Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, and Michael J. Stern. «Comparing Check-All and Forced-Choice Question Formats in Web Surveys». *Public Opinion Quarterly* 70, no.1(2006):66–77.
- [33] Lewis, James R., and Jeff Sauro. «Comparis on of Select-All-That-Apply Items with Yes/No Forced Choice Items». *Journal of Usability Studies* 17, no. 1 (2021):21–30.
- [34] Neuert, Cornelia E. «Do Forced-Choice (FC) Questions Trigger Deeper Cognitionthan Check-All-That-Apply (CATA) Questions?».*Journal of Survey Statistics and Methodology* 8, no. 4(2020):617–35.
- [36] Alwin, Duane F., and Brett A. Beattie. «The KISS Principle in Survey Design: Question Length and Data Quality». *SociologicalMethodology*46,no.1(2016):121–52.
- [37] Metzler, Anke, Tanja Kunz, and Marek Fuchs. «The Useand Position in gof Clarification Featuresin Web Surveys». *Psihologija* 48, no.4 (2015):379–408.
- [38] Peytchev, Andy.«Survey Breakoff». *Public Opinion Quarterly*, 73,no.1(2009):74–97.
- [39] Villar, Ana, Mario Callegaro, and Yongwei Yang. «Where am I? A Meta-Analysis of Experiments on the Effects of Progress Indicators for Web Surveys». *SocialScienceComputer Review*31,no.6(2013):744–62.
- [40] Lor, Maichou, Barbara J. Bowers, Anna Krupp, and Nora Jacobson. «Tailored Explanation: A Strategy to Minimize Nonresponse in Demographic Items Among Low-Income Racial and Ethnic Minorities». *Survey Practice* 10, no. 3. <https://doi.org/10.29115/SP-2017-0015>
- [41] Stats NZ. *2018 Census: Design of Forms*. Wellington: Stats NZ, 2018. <https://www.stats.govt.nz/assets/Reports/2018-census-design-of-forms/2018-Census-Design-of-forms.pdf>
- [42]Nichols, Elizabeth, Erica Olmsted-Hawala, Temika Holland , and Amy Anderson Riemer. «Usability Testing. Online Questionnaires: Experiences at the U.S. Census Bureau ». In *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by Paul Beatty, Debbie Collins, Lyn Kaye, Jose Luis Padilla, Gordon Willis, and Amanda Wilmot, 315–48. Hoboken: JohnWiley, 2020.
- [43] Chin, JohnP., Virginia A. Diehl, and Kent L. Norman. «Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface». In *CHI '88: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, edited by J. J.O'Hare, 213–18. New York: AssociationforComputingMachinery, 1988.
- [44] Rao, Gautam, and David Mingay. *Report on Usability Testing of Census Bureau's Dynamaps CD-ROM Product. 2001*. <https://infovis.cs.vt.edu/oldsite/cs5764/papers/dynamapsUsability.pdf>
- [45] Ericsson, K. Anders, and Herbert A. Simon. *Protocol analysis. Verbal Reportsas Data . Revised Edition*. Cambridge: MITPress, 1993.
- [46] Berti, Elisabetta. «Voci dal passato. La nostra memoria incisa su un nastro». *La Repubblica Firenze*, February, 11, 2023.

# Le mafie e le Digital Humanities

Marcello Ravveduto<sup>1</sup>, Camilla Zucchi<sup>2</sup>

<sup>1</sup> Università di Salerno, Italia – mravveduto@unisa.it

<sup>2</sup> Università di Salerno, Italia. – czucchi@unisa.it

## ABSTRACT

Lo studio nasce dalla impellente necessità di colmare una lacuna nei *Mafia Studies*: l'assenza di ricerche nell'ambito delle *Digital Humanities* o, più specificamente, un lavoro che integrasse il tradizionale approccio delle scienze umane con la *computer science* attraverso il *Web approach*. La squadra di ricerca è, proprio per questo, l'esito dell'unione fra competenze informatiche e umanistiche. Con l'intenzione di indagare l'immaginario mafioso veicolato sia su Wikipedia sia sui social, si è partiti dall'analisi delle voci presenti in Wikipedia, considerata in quanto comunità virtuale di scrittura collaborativa, su fatti e personaggi di mafia, per poi passare all'estrazione e all'interpretazione dei contenuti presenti sulle piattaforme social: YouTube, Twitter, Facebook, Instagram e TikTok.

## PAROLE CHIAVE

Digital Humanities; Mafia Studies; Social Network; Wikipedia; Digital Public History

## 1. LE MAFIE E LE DIGITAL HUMANITIES

«Lo storico di domani o sarà un programmatore o non sarà»<sup>1</sup>

Emmanuel Leroy Ladurie

Che cos'è l'immaginario? La similitudine più calzante per descriverlo è l'energia: non si crea né si distrugge, ma si trasforma continuamente. L'immaginario, come la parola suggerisce, è la percezione mentale, connessa alle esperienze personali, di un determinato fenomeno, che si produce attraverso i media. Tale processo è all'origine della formazione dei significati e delle risemantizzazioni delle parole e della realtà intera. Per fare qualche esempio, l'immaginario legato alla mafia, alla malavita e alle azioni di contrasto a queste ultime, è costituito da elementi eterogenei: la famosa foto di Falcone e Borsellino di Tony Gentile scattata il 27 marzo 1992 ha rappresentato, e rappresenta, un'icona della lotta alla criminalità organizzata; altrettanto si può dire per la saga de *Il padrino*, cui si ispirano, soprattutto all'estero, locali del *made in Italy* e contenuti dei social network; lo stesso vale per le canzoni di Niko Pandetta, trapper che nei suoi brani richiama lo stile di vita e i valori della criminalità organizzata<sup>2</sup>. Nessuno può sottrarsi al flusso persistente e pervasivo di immagini, suoni e parole che ogni giorno introiettiamo, anche passivamente. L'approdo al digitale si presenta come amplificazione del fenomeno: difatti, tale spazio si configura, da una parte, come lo specchio, dall'altra, come il motore dell'immaginario. Protagoniste di questa amplificazione sono le i millennial e la generazione z, che usano senza sforzo le varie risorse messe a disposizione dalla rete.

La ricerca è un focus mirato sull'immaginario digitale delle mafie. Lo studio offre un'efficace base di dati di partenza che ha permesso di monitorare le pratiche e gli 'spostamenti' delle associazioni criminali nell'uso delle piattaforme virtuali. Va sottolineato come la prassi di tali forme di presenza, nonché la produzione di significati orientati al sostegno e alla vicinanza a individui, non si sostanziano in riferimenti espliciti: si ricorre, frequentemente, a un codice non verbale, fondato, per esempio, sullo scambio di segni grafici e di immagini dotati di pregnanza simbolica in grado di esibire stili di vita emulativi, quali l'ostentazione della ricchezza e dello sfarzo al limite del kitsch e canzoni a sfondo criminale legate al contesto di appartenenza. Per comprendere i processi analizzati, ma anche quelli emersi dallo studio, è utile inserire la nozione di *mafie subdigitali*: organizzazioni criminali che agiscono e proliferano nell'ambiente digitale, facendo leva, in maniera indiretta, su stimoli visivi, sonori e comportamentali familiari a chi usa i social. Simili impulsi, nella maggior parte dei casi, sono troppo deboli per essere avvertiti a livello conscio, ma sono sufficienti a influenzare e condizionare la mentalità di alcuni individui. Si esalta, in questo modo, il ruolo dei media come veicoli, amplificatori e diffusori dei messaggi in codice. Così, crescendo il numero di followers, aumentano i potenziali emulativi e si diventa virali e "famosi". È la dimensione dell'interreale in cui virtuale e reale si influenzano a vicenda.

<sup>1</sup> «L'historien de demain sera programmeur ou il ne sera plus», *Le Territoire de l'historien*, (Paris: Gallimard, 1973), t. I.

<sup>2</sup> Canale Niko Pandetta <https://www.youtube.com/channel/UCDuX5tSzDq5MocnAJWQMxtA>, consultato il 3 maggio 2023.

L'estrazione dei dati rilevanti è stata effettuata con un duplice approccio: sono stati estratti manualmente 55 tra profili pagine e gruppi Facebook, 34 profili Instagram e 30 pagine Wikipedia, mentre automaticamente sono stati estratti 90 GB (11mila e 500 video di TikTok), 2 milioni e mezzo di tweet e 20mila commenti ai video YouTube.

## 2. L'ANALISI DI WIKIPEDIA

L'enciclopedia più usata è Wikipedia: la possibilità di accedere alla rete, la sua semplicità di consultazione e la necessità di avere le informazioni nel minor tempo possibile la rendono lo strumento privilegiato e più comodo<sup>3</sup>. Spesso si ricorre ad essa anche solo per avere un quadro d'insieme su una specifica questione o per dirimere un dubbio: il primo input è sempre quello di googlare il termine che vogliamo chiarire e successivamente aprire la relativa pagina Wikipedia. L'enciclopedia è tra i siti meglio ottimizzati, cosa che ne facilita la fruizione per gli utenti che si fermano alla prima SERP (*Search engine result page*) nel percorso di ricerca di informazioni. Se, dunque, non si contano i primissimi risultati, la *Serp heat map di informational queries* inerenti ad argomenti storici e culturali è monopolizzata da Wikipedia. Il web 2.0 ha dato la possibilità agli utenti di essere allo stesso tempo producer e consumer, cioè "prosumer"<sup>4</sup>: a creare e scrivere le pagine di Wikipedia sono, infatti, gli utenti grazie al modello wiki<sup>5</sup>. In questo modo, chi partecipa della scrittura di Wikipedia, può plasmare e riversare in essa il senso comune, maneggiando fonti secondarie. Inoltre, non conta l'identità dell'autore: si possono apportare modifiche diventando utenti editori o registrando il proprio indirizzo IP oppure si può solo visualizzare e rimanere utente lettore. Sono cinque i pilastri cui è richiesto di attenersi<sup>6</sup>:

- verificabilità: ogni informazione deve essere già presente in una fonte edita;
- neutralità: Wikipedia rimane imparziale, riportando ogni versione o teoria relativa a un argomento in base alla sua validità scientifica e alla validità delle fonti a suo
- sostegno;
- libertà: chiunque, rispettando le regole, può inserire o cambiare una voce;
- codice di condotta: Wikipedia ha un codice di condotta che va rispettato per evitare battaglie interpretative;
- non ci sono altre regole oltre a quelle elencate.

La presenza di amministratori dovrebbe garantire una mediazione pacifica in grado di appianare le discussioni. Su argomenti particolarmente sensibili, possono scatenarsi delle vere e proprie *edit wars*, specchio dei diversi punti di vista presenti nell'opinione pubblica. Wikipedia, dunque, è un luogo privilegiato per osservare il fluire dell'opinione pubblica e quali sono i temi "caldi", proprio perché più volte pubblicati e discussi. Tali premesse sono necessarie per comprendere il percorso di analisi effettuato. Un primo passo di analisi è stato quello di utilizzare una pagina web del 2007 contenente una cronologia (1893-2007) costituita dagli episodi criminali legati agli ambienti mafiosi considerati più rilevanti, stilata dalla Commissione Parlamentare Antimafia<sup>7</sup>. Si è, poi, verificato il grado di aderenza di Wikipedia alla suddetta cronologia: cioè se gli eventi e i personaggi protagonisti del documento fossero anche in Wikipedia. I riferimenti nella cronologia sono 310, su Wikipedia ne sono riscontrabili 207, con una percentuale di 66,7%. Successivamente, sono state individuate quattro categorie in cui suddividere gli argomenti prevalenti:

- boss, affiliati e collaboratori di giustizia;
- organizzazioni, clan e famiglie mafiose;

---

<sup>3</sup> Cfr. Gotor, Miguel "L'Isola Wikipedia. Una fonte elettronica" in Prima lezione di metodo storico, a cura di Sergio Luzzatto, 183-202, Roma-Bari: Laterza, 2010; Baldo, Tommaso. "Riflessioni sulla narrazione storica nelle voci di Wikipedia" *Diacronie* 29 (2017/1) [http://www.studistorici.com/2017/03/29/baldo\\_numero\\_29/](http://www.studistorici.com/2017/03/29/baldo_numero_29/); Baricco, Alessandro. *The game*, Torino: Einaudi, 2018 Pizzirusso, Igor. "Wikipedia: uno spazio per la Public History?" in *La storia liberata*, a cura di Paolo Bertella Farnetti e Cecilia Dau Novelli, 197-216, Milano-Udine: Mimesis edizioni, 2020. Wikipedia, voce Wikipedia Cinque Pilastri [https://it.wikipedia.org/wiki/Wikipedia:Cinque\\_pilastri](https://it.wikipedia.org/wiki/Wikipedia:Cinque_pilastri), consultato il 3 maggio 2023.

<sup>4</sup> Toffler, Alvin, *The Third Wave*, (New York: Morrow, 1980), 27.

<sup>5</sup> Paci, Deborah. "Conoscere è partecipare: digital public history, wiki e citizen humanities" *Umanistica Digitale* 10 (2021) DOI: <http://doi.org/10.6092/issn.2532-8816/12555>, consultato il 3 maggio 2023.

<sup>6</sup> Manera, Enrico. "Wikipedia come storia pubblica. Com'è e come potrebbe essere" *Novecento.org*, 22 marzo 2018 <https://www.novecento.org/uso-pubblico-della-storia/wikipedia-come-storia-pubblica-come-e-come-potrebbe-essere-3082/>, consultato il 3 maggio 2023.

<sup>7</sup> [https://web.archive.org/web/20071214175116/https://www.camera.it/\\_bicamerale/leg15/commbicantimafia/cronologiamafieantimafia/schedabase.asp](https://web.archive.org/web/20071214175116/https://www.camera.it/_bicamerale/leg15/commbicantimafia/cronologiamafieantimafia/schedabase.asp), consultato il 3 maggio 2023.

- stragi e vittime;
- personaggi coinvolti in attività mafiose o presunti tali.

Per la prima, la corrispondenza è del 57,6%, per la seconda 82%, per la terza 81%, per la quarta del 50%. Gli aspetti considerati indicatori d'interesse all'interno delle voci Wikipedia sono: la lunghezza e l'articolazione della pagina; la presenza della sezione Discussione o del paragrafo dedicato all'influenza del personaggio, o evento, nella cultura di massa; l'azione degli utenti, sia lettori sia editori in termini di visualizzazioni e interventi diretti<sup>8</sup>; l'estensione di Note e Bibliografia. In queste due sezioni la stampa è la fonte dominante sia per la categoria boss, affiliati e collaboratori di giustizia, sia per quella stragi e vittime di mafia.

Prendendo come metro di paragone l'indice del progetto Vivi – Nomi da non dimenticare, realizzato dall'associazione Libera<sup>9</sup>, risulta nettamente più bassa, invece, la corrispondenza tra l'elenco di vittime innocenti del portale e quelle presenti in Wikipedia: circa il 19%.

Per dare l'idea di quali siano le voci maggiormente oggetto di revisioni nel 2022 rispetto al 2021, sono quelle di Michele Greco, Piersanti Mattarella e Salvatore Contorno, mentre diminuiscono quelle relative a Giovanni Brusca, Salvatore Riina e Benedetto Santapaola.

Le pagine più elaborate si confermano essere specchio e motore del senso comune. Infatti, le modifiche riguardano i criminali più noti e delle vittime più ricorrenti nella memoria pubblica e avvengono soprattutto in corrispondenza degli anniversari. Quando si è di fronte a modifiche frequenti, si può affermare che vi è ancora un interesse vivo per un determinato evento o personaggio; se, invece, le visualizzazioni sono negli anni costantemente alte ma con un numero basso di interventi, si può pensare ad una narrazione condivisa e pacifica.

### 3. L'ANALISI DEI SOCIAL NETWORK

I social network sono il grande palcoscenico mediatico degli ultimi quindici anni: permettono l'interazione immediata attraverso post, reazioni, commenti, storie, dirette live e messaggi privati istantanei, e, quanto all'oggetto della ricerca, restituiscono un'immagine molto aderente sia di ciò che l'opinione pubblica pensa del fenomeno mafioso sia di ciò che viene condiviso da parte di affiliati e simpatizzanti. La proliferazione di piattaforme *ugc* ha comportato la progressiva specializzazione dei diversi media. YouTube è veicolo di programmi tv e video musicali, Facebook di produzione di contenuti individuali e di gruppi privati, Twitter prevalentemente di comunicazioni ufficiali e istituzionali, Instagram è stato l'incubatore di *influencer* e *content creator* e oggi è una piattaforma *glamour* in cui si ostenta benessere e si pubblicizzano beni di consumo. Quanto al più recente TikTok, si presenta ormai come il centro nevralgico del nuovo modo di comunicare: utilizza tracce sonore, risemantizzate dall'utente, per costruire video in grado di raggiungere una diffusione virale. Naturalmente, nel loro insieme hanno cambiato il modo di relazionarsi delle persone e la loro visione del mondo. Le organizzazioni criminali utilizzano i social network per comunicare, affermarsi ed espandersi, adeguandosi a questo nuovo palcoscenico. Le tracce sonore combinate ai contenuti, ad esempio, non sono mai casuali. Allo stesso modo, gli emoji non sono solo corredo del testo, ma, spesso, propongono una lettura semantica pluricodice nota a mittente e destinatari. Attorno al leader che occupa il centro dei contatti, si costituisce una cerchia di individui che mostrano di essere simpatizzanti. Questa analisi è particolarmente utile e significativa in quanto permette di ricostruire il sistema di relazioni del network criminale. Pertanto, i social network rivestono un ruolo strategico per monitorare le attività dei mafiosi e della loro cerchia, che sono spesso il riflesso, il prosieguo o la risposta alla realtà.

#### YouTube

YouTube è l'esempio più rilevante e significativo di archivio transmediale, oltre ad essere una piattaforma di condivisione che consente anche l'interazione immediata attraverso commenti e like/dislike. Al suo interno vengono caricati sia video musicali sia video divulgativi, tratti da programmi tv o montati da appassionati. Ciò lo rende assimilabile ad un mezzo di informazione. Si è scelto di partire da un campione di cento video per entrambe le tipologie sopra citate, utilizzando parole chiave e l'algoritmo della piattaforma che suggerisce contenuti affini.

Con riferimento ai video di tipo divulgativo, il campione finale casuale, spinto dall'algoritmo, cresce a 265 video. I contenuti analizzati provengono da varie fonti sia pubbliche sia private. Con il software ExportComments<sup>10</sup>, sono stati

<sup>8</sup> Strumento usato:

<https://pageviews.wmcloud.org/?project=it.wikipedia.org&platform=allaccess&agent=user&redirects=0&start=2021-01&end=2021-11&pages=>, consultato il 3 maggio 2023.

<sup>9</sup> <https://vivi.libera.it/>, consultato il 3 maggio 2023.

<sup>10</sup> <https://exportcomments.com/>, consultato il 3 maggio 2023.

estratti e analizzati più di 10mila commenti, il cui *sentiment*, intendendo l'accezione semantica correlata al contenuto, è eticamente marcato in maniera negativa nei confronti delle mafie.

I video musicali analizzati riguardano, invece, canzoni che inneggiano alla malavita, agli "uomini d'onore", disprezzando "sbirri e infami". Si inseriscono quasi tutti nei generi della trap e del neomelodico, spesso associati ad ambienti mafiosi. I video, corredati da elementi visivi e stilistici ricorrenti, veicolano un immaginario che si ritrova anche negli altri social: abbigliamento *streetwear* di lusso, gioielli vistosi e pieni di oro e diamanti, macchine costose e vita notturna sfrenata. I commenti analizzati automaticamente, relativi ai video musicali, hanno un *sentiment* positivo e di incoraggiamento all'artista e sono spesso scritti in dialetto.

### Facebook

Più volte dato per spacciato a causa dalla fama crescente dei concorrenti, Facebook è il più longevo dei social network. Il suo utilizzo è stato di immediata comprensione. Si possono scrivere post o pubblicare foto in modo facile e intuitivo, commentare quelli degli amici e inviarsi messaggi privati, con un processo di interazione elementare.

Riguardo la nostra analisi, si evidenzia, accanto all'esistenza di profili personali, la presenza di pagine e gruppi, facili da esplorare quando non impostati con limitazioni della privacy. Le pagine che hanno destato maggiore attenzione sono quelle dei fan di cantanti trap e neomelodici come Niko Pandetta e Daniele de Martino oppure quelle di personaggi pubblici che denunciano la malavita, come Francesco Emilio Borrelli, Giovanni Russo, Josi Della Ragione. Dai contenuti condivisi da questi ultimi, si sono potute analizzare le reazioni nei commenti, che si alternano tra disprezzo degli episodi criminosi e inviti reiterati all'autore all'indifferenza omertosa. Dall'osservazione dei gruppi, invece, è emerso che la maggioranza ha uno status di privato e, quindi, inaccessibile, se non per informazioni di base. Sono stati osservati undici gruppi, che hanno con un range tra gli 8 e 17mila membri. Fra i più significativi, è sicuramente quello dedicato al già citato cantante trap-neomelodico Niko Pandetta, il gruppo "Niko Pandetta Gruppo Ufficiale", con 809 iscritti.

### Twitter

La natura rapida dei tweet, che prevedono un massimo di 280 caratteri, fa della piattaforma il luogo in cui ci si informa in tempo reale seguendo *trend topic*. Sono quasi 4 milioni gli italiani iscritti a Twitter. Nel corso del tempo si è caratterizzato come il social degli *opinion maker* per commentare fatti ed eventi.

Ai fini della ricerca, sono stati estratti 2 milioni e mezzo di tweet attraverso alcune parole chiave in grado di restituirci il pensiero dell'opinione pubblica su episodi o personaggi legati alle organizzazioni mafiose. È opportuno notare che la decrescita dei tweet riferibili al tema della ricerca non è associabile al minor utilizzo della piattaforma: infatti, nell'anno 2020 assistiamo, da un lato, a un drastico calo di contenuti relativi alle mafie, dall'altro, all'apice nell'uso della piattaforma. Molto probabilmente questo fenomeno è legato alla pandemia da SARS-CoV-2. Dal 2012 ad oggi si nota che tra i dieci utenti più produttivi in termini di tweet sono @concentrazione, per Concentrazione Antimafia, e @infoiterno per Informazione interno. Nell'arco temporale si sono analizzati anche i messaggi più ritwittati, i cui autori sono per lo più personaggi noti come Roberto Saviano, Erri De Luca, Matteo Renzi, Laura Boldrini o testate giornalistiche, per esempio *la Repubblica*. Successivamente, accorpando il testo dei tweet di anno in anno, si è effettuato il calcolo delle parole più frequenti: spiccano con continuità "ndrangheta", "camorra", "mafia" e "boss". Le prime tre sono già contenute nella query di partenza. Va, tuttavia, rilevato che la parola mafia, spesso, si comporta come un ombrello semantico.

### Instagram

Confluito anch'esso da qualche anno nel gruppo Meta, Instagram rappresenta il luogo ideale dove condividere le proprie foto con una breve descrizione a volte accompagnata da hashtag. Dopo l'esplosione di TikTok, è stato reso possibile anche su Instagram girare e montare video performance, chiamati reel. Di natura completamente diversa rispetto a Facebook, Instagram è il primo social network pensato esclusivamente per *mobile*. Con la sua comparsa, si è affermata la professione di *influencer* e *content creator*, che realizzano contenuti ad hoc per promuovere prodotti alla luce della loro popolarità. Anche i mafiosi agiscono e si offrono grazie a questa piattaforma allo sguardo di utenti e follower mostrandosi in atteggiamenti *glamour* con oggetti di valore, automobili costose, accessori di lussi e vestiti esclusivi della linea *streeturban*. Quasi sempre, i loro profili sono privati e consentono l'interazione solo a persone fidate. Sottolineiamo che in questo caso troviamo una spasmodica attenzione alla privacy poiché alcuni affiliati che hanno un profilo privato su Instagram, mantengono un profilo pubblico su altri social.

### TikTok



TikTok è nato nel 2016, è la più giovane tra le piattaforme analizzate ed è anche la più interessante. I contenuti sono video che hanno una durata da un minimo di 15 secondi a un massimo di 10 minuti. Possono essere provvisti di una rapida descrizione e consentono di raggiungere una fama effimera a chi li produce se compaiono nella sezione dei “Per te” diversa per ciascun utente. Gli autori dei video che diventano virali acquisiscono lo status di *content creator* e possono monetizzare il proprio successo. Più video si producono, più *trend* si seguono, più si ha la possibilità di diventare virali. I video possono essere dei *mash up* con tracce sonore e immagini risemantizzate oppure performance originali. Per la nostra ricerca, sono stati selezionati, attraverso gli hashtag, 11 mila e 508 video ed elaborati 70 mila kiloByte di dati. L’impennata dei contenuti correlati ai nostri hashtag inizia nel 2020, forse favorita dalla pandemia, e continua ancora nel 2022, soprattutto nel mese di luglio. Molto probabilmente, questo dato può essere messo in relazione alla frequenza di hashtag relativi ai giudici Giovanni Falcone e Paolo Borsellino (*#falcone*, *#borsellino*), in funzione del trentennale della strage di via D’Amelio. L’hashtag *#mafia*, come già detto, ha funzione di ombrello semantico: raccoglie, infatti, contenuti di personaggi verosimilmente legati alle organizzazioni mafiose ma anche montaggi con foto o spezzoni di film divenuti parte dell’immaginario collettivo; per esempio, *Scarface*, *Il padrino*, *Gomorra*. Altri hashtag rimandano alla vita in carcere o ai domiciliari e non di rado sono accompagnati da canzoni neomelodiche o da emoji che hanno un preciso destinatario in grado di coglierne il significato recondito. Il cuore, le catene e il leone sono gli emoji più frequenti e vengono utilizzati spesso per comunicare rispettivamente affetto, detenzione e orgoglio. Le tracce musicali che accompagnano questi TikTok, dando un significato preciso al video, attingono a piene mani all’universo neomelodico e trap dialettale: si va, ad esempio, da “Rispetto ‘e libertà” di Nello Amato e “Nu carcerato” a “Pistole nella Fendi” di Niko Pandetta, che, come si evince dal titolo, mescola vita criminale (pistole) e lusso (Fendi). Non marginali sono il ruolo e l’attività dei profili di coppia, spesso gestiti dalla compagna del carcerato che, con estrema devozione, gli dedica video e canzoni in attesa del ritorno.

In linea generale, emergono alcuni filoni predominanti nei video TikTok: quello commemorativo per gli affiliati morti – interessante il caso di Emanuele Sibillo –, quello celebrativo per chi è in carcere o ha un ruolo apicale nel clan, infine, quello che può essere definito propagandistico e denigratorio dei rivali. Nel loro insieme, questi filoni sono portati avanti dai simpatizzanti della cerchia criminale. Si riscontrano, dunque, delle linee di tendenza: i video spesso ritraggono elementi di lusso e ricchezza o anche tatuaggi con richiami specifici; si parla spesso in dialetto con un sottofondo di canzoni neomelodiche e trap; alcuni sono accompagnati da emoji e rappresentano scene con riferimenti al carcere e/o all’arresto e si prestano a diventare occasioni di commemorazione. TikTok si offre, perciò, insieme ad Instagram, come piattaforma di marketing orientata alla promozione/celebrazione di un certo stile di vita associabile all’immaginario delle mafie.

Tra le varie organizzazioni criminali, la camorra, ad oggi, sembra essere la maggiore produttrice di contenuti sulla piattaforma. Tuttavia, molti elementi trascendono l’appartenenza al singolo clan: ad esempio, il lusso e il ricorso a musiche neomelodiche sono comuni anche nei contenuti che si avvicinano di più alla malavita romana dei Casamonica. Superano, invece, la barriera linguistica, diventando così un codice transnazionale, sia gli emoji sia il lusso-kitsch, elemento in comune con i narcotrafficanti sudamericani.

## BIBLIOGRAFIA

- [1] Baldo, Tommaso. “Riflessioni sulla narrazione storica nelle voci di Wikipedia” *Diacronie* 29 (2017/1) [http://www.studistorici.com/2017/03/29/baldo\\_numero\\_29/](http://www.studistorici.com/2017/03/29/baldo_numero_29/)
- [2] Baricco, Alessandro. *The game*, Torino: Einaudi, 2018
- [3] Commissione parlamentare antimafia, Cronologia su mafia e antimafia, [https://web.archive.org/web/20071214175116/https://www.camera.it/\\_bicamerale/leg15/commbicantimafia/cronologiamafieantimafia/schedabase.asp](https://web.archive.org/web/20071214175116/https://www.camera.it/_bicamerale/leg15/commbicantimafia/cronologiamafieantimafia/schedabase.asp), consultato il 3 maggio 2023.
- [4] Gotor, Miguel “L’Isola Wikipedia. Una fonte elettronica” in *Prima lezione di metodo storico*, a cura di Sergio Luzzatto, 183-202, Roma-Bari: Laterza, 2010.
- [5] LeRoy Ladurie, Emmanuel. *Le Territoire de l’historien*, Gallimard, Paris, t. I, (1973)
- [6] Manera, Enrico. “Wikipedia come storia pubblica. Com’è e come potrebbe essere” *Novecento.org*, 22 marzo 2018 <https://www.novecento.org/uso-pubblico-della-storia/wikipedia-come-storia-pubblica-come-e-come-potrebbe-essere-3082/>, consultato il 3 maggio 2023.
- [7] Canale Niko Pandetta <https://www.youtube.com/channel/UCDuX5tSzDq5MocnAJWQMxTA>, consultato il 3 maggio 2023.
- [8] Paci, Deborah. “Conoscere è partecipare: digital public history, wiki e citizen humanities” *Umanistica Digitale* 10 (2021) DOI: <http://doi.org/10.6092/issn.2532-8816/12555>, consultato il 3 maggio 2023.
- [9] Pizzirusso, Igor. “Wikipedia: uno spazio per la Public History?” in *La storia liberata*, a cura di Paolo Bertella Farnetti e Cecilia Dau Novelli, 197-216, Milano-Udine: Mimesis edizioni, 2020.
- [10] Toffler, Alvin, *The Third Wave*, New York: Morrow, 1980.
- [11] Wikipedia, voce Wikipedia Cinque Pilastrini [https://it.wikipedia.org/wiki/Wikipedia:Cinque\\_pilastrini](https://it.wikipedia.org/wiki/Wikipedia:Cinque_pilastrini), consultato il 3 maggio 2023.



## **SITOGRAFIA**

<https://exportcomments.com/>

<https://pageviews.wmcloud.org/?project=it.wikipedia.org&platform=allaccess&agent=user&redirects=0&start=2021-01&end=2021-11&pages=>

<https://vivi.libera.it/>

# I Confini della Lunigiana: un progetto di Digital Public History

Enrica Salvatori<sup>1</sup>, Monica Bacci<sup>2</sup>

<sup>1</sup> Università di Pisa, Italia – enrica.salvatori@unipi.it

<sup>2</sup> Università di Pisa, Italia – monicabacci8@yahoo.it

## ABSTRACT

Il progetto di Digital Public History “I confini della Lunigiana” ha lo scopo di analizzare quanto il concetto di Lunigiana sia radicato nelle persone che attualmente la abitano, che sono originarie dello spazio tradizionalmente definito come lunigianese o che la frequentano assiduamente. Attraverso un questionario online gli intervistati hanno la possibilità di dare il loro contributo rivelando in che modo percepiscono la storia in questo territorio e quindi di riflettere sulle ragioni che contribuiscono al loro eventuale senso di appartenenza a questo spazio privo di confini politico-amministrativi.

## PAROLE CHIAVE

Lunigiana; Digital Public History; Identità; Cultural Economy; Storia pubblica digitale

## 1. INTRODUZIONE

Il progetto di ricerca *I confini della Lunigiana*, attivo dal 2022, si configura come un progetto di Storia Pubblica Digitale in quanto ha l'ambizione di andare a reperire e analizzare informazioni appartenenti al patrimonio culturale di quanti, a vario titolo, frequentano o hanno frequentato questa terra tramite strumenti appartenenti al dominio delle Digital Humanities. In particolare, si prefigge di analizzare in che modo i pubblici lunigianesi percepiscono la storia di questo territorio portandoli a riflettere sulle ragioni che contribuiscono o meno alla formazione dell'identità delle persone in uno spazio ricco di storia, ma dai confini politico-amministrativi non definiti.

La partecipazione del pubblico al questionario, promossa anche con eventi pubblici di presentazione e in seguito la possibilità di vedere in tempo reale i risultati, ha lo scopo, oltre che di reperire e analizzare le informazioni, di portare a una maggiore consapevolezza e un più ampio coinvolgimento dei diversi pubblici alla storia locale; ha inoltre reso l'utente un *prosumer* nel senso di cittadino consapevole e al tempo stesso promotore di storia, utile alla collettività in quanto viene chiamato a partecipare al progetto di ricerca, ai suoi quesiti, fino all'interpretazione dell'intera esperienza collettiva.

Da una prima fase di studio della Lunigiana e degli aspetti storici e antropologici di questa terra, si è passati alla costruzione di un questionario da compilare online con sua contestuale promozione in eventi pubblici e sui social. Essendo rivolto a un pubblico maggiore di 13 anni di età (terzo anno della scuola superiore di primo grado) abbiamo promosso la ricerca attraverso i canali social più frequentati da un pubblico adulto (Facebook, Instagram) attuando contemporaneamente una campagna capillare via mail, indirizzata ad associazioni locali, Comuni, quotidiani locali, musei del territorio e istituti scolastici.

La raccolta dei dati, la successiva elaborazione degli stessi e la loro restituzione si è fondata su scelte tecniche di uso comune, che consentissero una relativa facilità d'uso, in particolare:

- G Suite, la nota suite di software e strumenti di produttività per il cloud computing e per la collaborazione, offerta da Google: in particolare Google Moduli, inclusa nella suite per ufficio di Google Drive. La raccolta dei dati è avvenuta tramite la creazione di una pagina del questionario impostata con un'immagine iniziale dell'Archivio di Stato di Firenze di una carta antica della Lunigiana risalente al 1767. L'immagine contenente il testo sul trattamento dei dati soggetti alla privacy, inserita all'inizio del questionario, non rispettando i requisiti di accessibilità delle pagine web, è stata tradotta in testo OCR attraverso l'utilizzo dell'app Keep compresa in G Suite. In oltre nel questionario è stato necessario inserire un'espressione regolare per il controllo degli indirizzi e-mail:  $^[A-z0-9\.\+\_]+\@[A-z0-9\.\+\_]+\.[a-z]{2,6}$ .
- La restituzione dei dati, essendo la raccolta dilatata nel tempo, ha previsto un canale di visualizzazione web in cui i dati possono essere visualizzati in tempo reale in modalità interattiva, ossia con l'opportunità per l'utente di selezionare eventuali filtri di suo gradimento e ottenere immediatamente l'elaborazione dei dati

conseguenziale alla scelta richiesta. Questa possibilità si è resa possibile attraverso l'installazione del componente aggiuntivo Awesome Table (Fig. 1) per fogli di calcolo Google. Una volta installato, l'applicativo crea una app attraverso la quale qualsiasi utente può accedere all'elaborazione grafica dei risultati dell'indagine senza alcuna registrazione. Awesome Table, infatti, come ultima istanza dell'installazione, genera un link associato al foglio di calcolo e non solo produce i grafici dai dati raccolti, ma li rende interattivi direttamente dall'utente ricalcolando i risultati, aggiornando i grafici, subito dopo la selezione operata dall'utente. La app genera inoltre uno script affinché la visualizzazione interattiva dei risultati del questionario possa essere inserita in una pagina web.

- Particolari elaborazioni, per uno studio più approfondito, sono state affrontate con Notebook Jupyter, un'applicazione basata sul modello client-server dell'organizzazione no-profit Progetto Jupyter, fondata nel 2015 e disponibile gratuitamente con licenza BSD.
- L'elaborazione dei dati geografici relativi ai Comuni più frequentati e maggiormente identitari della Lunigiana, sono confluiti in un GIS (Geographic Information System) la cui mappa è stata inserita nel poster.
- Il lavoro di programmazione sui dati raccolti è stato orientato più alla produzione di grafici piuttosto che di tabelle perché ritenuta una visualizzazione dati efficace alla comprensione immediata del fenomeno. In seguito, a progetto concluso, i dati verranno rilasciati anche in altri formati. Per la rappresentazione grafica dei dati, non essendo questi di tipo numerico per la quasi totalità delle variabili, sono stati scelti prevalentemente grafici di tipo Line Charts (su valori contigui), Pie Chart (per enfatizzare le differenze in proporzione), Bar Graph (per confronti di valori relativi) e box Plot in alcuni casi per fare comparazioni di statistica descrittiva.



Figura1. App Awesome Table inserita nel sito web.

Attualmente il questionario è stato compilato da circa 550 persone che, data la scarsità demografica della regione e la scarsa abitudine a usare la rete, specialmente per le persone anziane, è un dato accettabile, sebbene non completamente soddisfacente. I risultati sono stati comunque interessanti: è emerso un gruppo umano compatto nel ritenere che l'attaccamento a questa terra non riguardi questioni amministrativo-politiche, ma sia da ricercare su fattori di lunga durata, primo fra tutti il paesaggio antropizzato. Emerge però anche il timore della perdita dell'identità culturale.

Gli intervistati ritengono, infatti, sia importante svolgere a scuola attività finalizzate alla scoperta del passato lunigianese, alla promozione dello sviluppo economico e turistico, alla creazione di un'economia dei piccoli borghi e, nel lungo termine, prevedono che l'istituzione di una nuova Provincia o Regione, possa portare benefici sul piano del ripopolamento di territori abbandonati e sulla promozione dello sviluppo economico.

In questo contesto le naturali conseguenze di tutta questa ricerca porterebbero a suggerire azioni nel settore della *cultural economy* intesa come l'insieme degli studi sulle risorse culturali e sulle attività a esse correlate al fine di valorizzare le economie locali. La *cultural economy* fornisce una visione "economica" agli studi culturali e dovrebbe favorire la presa di coscienza del patrimonio culturale di un territorio per la sua valorizzazione.

Sito del progetto con link al questionario. <http://www.labcd.unipi.it/progetti/i-confini-della-lunigiana/>

Risultati in tempo reale. <https://sites.google.com/view/confinilunigiana/home-page>

Link al poster. <https://sites.google.com/view/confinilunigiana/home-page/poster>

## **BIBLIOGRAFIA**

- [1] Dear, Michael, The Postmodern Turn, in Minca Claudio (a cura di), *Postmodern Geography. Theory and praxis*, Malden, Blackwell, 2001.
- [2] Dei, Fabio, *Antropologia culturale*, Il Mulino, Bologna, 2016.
- [3] Gallo, Nicola. *Cartografia storica e territorio della Lunigiana centro orientale*. Sarzana, Lunaria, 1993.
- [4] Rosenzweig, Roy. Can History Be Open Source? Wikipedia and the Future of the Past, in «*The Journal of American History*», XCIII (2006), n. 1, pp. 117–146.
- [5] Mazzini, Ubaldo. Per i confini della Lunigiana. in «*Giornale storico della Lunigiana*», I (1909), pp. 4-38.
- [6] Pistarino, Geo., La Lunigiana storica, in «*Memorie dell'Accademia Lunigianese di Scienze e Lettere 'G. Capellini'*», LIV–LVI (1984-1986), pp. 3–22.
- [7] Salvatori, Enrica. Fosdinovo nel quadro delle gerarchie delle strade e dei poteri della Lunigiana medievale, in «*Memorie dell'accademia lunigianese di Scienze 'Giovanni Capellini'*», LXXII (2002), pp. 39-56;
- [8] Salvatori, Enrica, Digital Public History Inside and Outside the box, in «*Magazén. International Journal for Digital and Public Humanities*», 1/2 (2020), pp. 203-222.

# L'immaginario francese della Puglia: una collezione digitale del 'Petit Tour' dell'altra Italia

Michele Sollecito<sup>1</sup>

<sup>1</sup> Università degli studi di Bari Aldo Moro, Italia – michele.sollecito@uniba.it

## ABSTRACT

La presente proposta si incentra sull'asse quattro della Call for paper, ovvero sulla promozione delle buone pratiche che garantiscono preservazione e accesso al patrimonio digitale. In virtù del programma della Regione Puglia, "Research for Innovation" (REFIN), è stato possibile finanziare progetti di ricerca per la crescita intelligente, inclusiva e sostenibile del territorio. Il contributo qui presente si occupa di fornire un excursus su questa iniziativa che ha permesso la digitalizzazione dei testi sull'immaginario francese della Puglia al fine di poter curare e promuovere al meglio l'immagine della regione destinata al pubblico francese, un mercato turistico che per la Puglia costituisce il secondo mercato globale in assoluto.

## PAROLE CHIAVE

Digitalizzazione; Puglia; Turismo; Mappe; Viaggiatori francesi

## 1. IL 'PETIT TOUR' DELL'ALTRA ITALIA

L'immaginario francese della Puglia ha contribuito non poco alla promozione regionale. Se oggi la Francia costituisce il secondo mercato del turismo straniero in Puglia lo si deve a una attenzione particolare da parte dei viaggiatori francesi verso il "tacco dell'Italia". Mentre nel tempo – tra il Rinascimento e il Risorgimento italiano – si consolidava in Italia lo statuto del "Grand Tour" con le tappe obbligate tra Nord e Centro, parimenti si faceva strada il racconto del Sud Italia a cura di viaggiatori meno illustri rispetto ai letterati che frequentarono l'"altra" Italia. Occhi stupiti dettavano reportage ameni, restituivano contesti favolistici e paesaggi quasi ancestrali. Questi testi costituirono la base ideale per la valorizzazione in chiave turistica della Puglia, valorizzazione che avvenne da parte dei francesi in modo deciso sin dai primi anni Sessanta del Novecento.

A partire dalla bibliografia già esistente sui viaggiatori francesi in Puglia dal Cinquecento al Novecento<sup>1</sup> si è proceduto alla digitalizzazione delle pubblicazioni significative e utili a ricostruire le categorie dell'immaginario francese della Puglia ottenendo un Corpus di quarantasei testi da Gilles Le Bouvier (XV secolo) a André Pieyre de Mandiagues (XX secolo).

Al termine del lavoro di digitalizzazione e opportuna verifica filologica dei testi è stato possibile marcare elettronicamente i personaggi storici nonché i luoghi citati e visitati da questi viaggiatori utilizzando la piattaforma Recogito in virtù dell'agilità d'uso e della notevole precisione dovuta ai Gazetteer disponibili per le marcature tra cui: HistoGIS<sup>2</sup>, Pleiades<sup>3</sup>, DARE<sup>4</sup>, GeoNames<sup>5</sup>. Questi file XML-TEI costituiranno così il database del sito web e webapp dell'immaginario francese della Puglia in corso d'opera. Nello specifico, il motore di consultazione del database renderà disponibili livelli di interrogazione di diversa tipologia e profondità permettendo ricerche per autore, per anno di edizione, per luoghi e persone citate in uno o più testi presenti nel database. Il sistema permetterà all'utente di navigare il contenuto anche per "Categorie dell'immaginario francese della Puglia" illustrando gli itinerari degli autori e gli itinerari dell'immaginario. La metadattazione delle opere, infatti, conterrà non solo informazioni editoriali ma anche informazioni sugli itinerari di viaggio a seconda delle categorie dell'immaginario prevalenti e peculiari<sup>6</sup>.

---

<sup>1</sup> Dotoli Giovanni, Fiorino Fulvia. *Viaggiatori francesi in Puglia*. 8 voll., Fasano: Schena, 1985-1999. Dal XV al XX secolo sono state digitalizzate quasi tutte le opere che il catalogo della Bibliothèque Nationale de France riporta come opere afferenti alla Puglia o al Mezzogiorno. Sono state escluse alcune opere specialistiche afferenti a questioni geomorfologiche, climatiche, geografiche, industriali mentre sono state preferite le opere che si concentravano maggiormente sulle impressioni di viaggio, sui costumi, sulla cultura, sulla storia, sul paesaggio. Lo spoglio analitico dei testi del secondo Novecento e dei primi anni del 2000, coperti da diritto d'autore, ha fornito interessanti spunti per l'apparato critico e per ricostruire il cambiamento della percezione della Puglia nei confronti dei viaggiatori e scrittori francesi soprattutto negli ultimi anni.

<sup>2</sup> A GIS repository for historical temporalized spatial data by the Austrian Centre for Digital Humanities.

<sup>3</sup> Pleiades Gazetteer of the Ancient World.

<sup>4</sup> Digital Atlas of the Roman Empire.

<sup>5</sup> A subset of GeoNames populated places, countries and first-level administrative divisions.

<sup>6</sup> Tutti i dati presenti sulla piattaforma saranno pubblicamente interrogabili e sarà possibile scaricarne il contenuto in formato XML o JSON.

Il sito e la web application, esito finale del progetto finanziato dall'iniziativa regionale REFIN, saranno oggetto di utilizzo da parte di PugliaPromozione (Agenzia regionale del Turismo – ARET) per l'organizzazione di incontri ed *educational tour* con operatori di settore interessati al mercato francese.

L'obiettivo è veicolare, tramite questi operatori, una proposta della Puglia che sia non solo legata alla fruizione del mare ma anche ai suoi beni culturali e paesaggisti, alle impressioni di viaggio dei tanti scrittori francesi che nel tempo hanno scoperto scorci di Puglia inaspettati, descritti in modo accattivante ed interessante. La georeferenziazione dei luoghi citati permette, infatti, di mappare i percorsi di questi viaggiatori e quindi di suggerire itinerari "letterari", itinerari dalla valenza storica, artistica, infine itinerari per categorie dell'immaginario francese della Puglia. Le mappe ottenute ci permettono di apprezzare meglio come nel corso del tempo, sino ad arrivare al giorno d'oggi, siano cambiate le attenzioni dei viaggiatori francesi per la Puglia pur restando immutato lo stupore per il valore paesaggistico di una regione che – come una costante nei diversi testi – resta un luogo dove, a tratti, tempo e spazio si arrestano e collocano il viaggiatore in un altrove esotico, pittoresco e arcaico. Come osserva Barbara Piatti, nella migliore tradizione letteraria degli studi cartografici la visualizzazione delle mappe non sostituisce i metodi ermeneutici classici bensì ci permette di tornare alle fonti primarie – ai testi – per illustrare ciò che è desumibile, ispirare ciò che altrimenti si dava per scontato. L'approccio filologico rivela un senso nuovo a ciò che invece finora era rimasto nascosto<sup>7</sup>.

## 2. LE CATEGORIE DELL'IMMAGINARIO FRANCESE DELLA PUGLIA NEI TESTI DAL CINQUECENTO AL NOVECENTO

Se è vero che scrittori che hanno lasciato importanti note di viaggio sull'Italia come Montaigne, Charles De Brosse, Montesquieu, Stendhal, i Goncourt, Gide, Sartre, hanno trascurato il Sud, ad eccezione di Napoli, è altrettanto vero che gli altri viaggiatori-scrittori che si sono avventurati a sud del capoluogo partenopeo hanno lasciato pagine di indiscutibile bellezza per profondità ed accuratezza della descrizione, per l'erudito apparato storico, per la meraviglia e lo stupore generatosi a fronte di un pregiudizio negativo sul Sud dimostratosi più volte infondato.

Diverse sono state le scoperte interessanti, tutte ascrivibili a categorie dell'immaginario che consente di scortare il lettore o il turista curioso alla ricerca di motivazioni per decidere di partire alla scoperta della Puglia.

La Puglia che ricaviamo da questi testi ha volti diversi, un fascino che si ricava dalla somma delle impressioni visive ed esperienziali che possono alimentare l'immaginazione secondo un motivo allegorico, architettonico, bucolico, esotico, gastronomico, mitologico, pittoresco, religioso, romantico, storico, umanistico-rinascimentale.

Questo caleidoscopio di immagini spazia dal richiamo dantesco "allegorico" dell'immagine della Chiesa – "Colei che è assisa sulle acque" – riferita da André Maurel alla splendida cattedrale di Trani<sup>8</sup>, sino al paragone "pittoresco" del Tavoliere con la Beauce francese, un'immensa pianura ingiallita dai raccolti maturi così come tratteggiata da Charles Yriarte. La Puglia rappresenta spesso un ritorno a un passato mitico, quasi arcaico, un quadro "bucolico" secondo le impressioni di Eugène Marsan : un mondo di pastori, di mandrie, prati, stalle, vomeri, di canti trasmessi dalla poesia greca e latina ormai dimenticati nei salotti<sup>9</sup>.

La Puglia è anche – e soprattutto – il suo mare, una scoperta singolare in quanto al di là dei suoi porti – Bari e Brindisi – passaggi obbligati per il viaggio in Oriente, vi è un mare dal valore intrinseco notevole. Tale fu la scoperta dell'Abate di Saint-Non dinanzi allo spettacolo delle grotte di Polignano, luoghi dove la frescura delle acque e il loro colore cangiante costituivano un'esperienza singolare e affascinante<sup>10</sup>. L'immaginario francese tocca anche le città pugliesi, si spazia facilmente dal richiamo "umanistico-rinascimentale" in riferimento a Lecce, capitale del barocco, definita spesso la "Firenze del Sud", sino all'immagine della ricchezza "gastronomica" di Foggia, città-deposito del grano del Tavoliere i cui magazzini sono approvvigionati di grasse mortadelle, salsicce, formaggi, conserve di ogni natura. Non manca, inoltre, la particolare attenzione all'arte romanica e in particolare alle cattedrali sul mare, da Barletta a Bari. La basilica di San Nicola è considerata la pura chiesa normanna in linea diretta con la cattedrale di Caen<sup>11</sup>, così come grande interesse "religioso"

<sup>7</sup> Piatti, Barbara. "Literary Cartography: Mapping as Methods", in Engberg-Pedersen Anders (ed.), *Literature and cartography: Theories, histories, genres*. 58-60, MIT Press, 2017.

<sup>8</sup> Maurel, André. *Petites villes d'Italie*. Vol. 3, Paris : Hachette.

<sup>9</sup> Marsan, Eugène. *Italie méridionale*, in *Le Visage de l'Italie*. Paris : Horizons de France, 1929, 270 : « Ce monde des bergers et des troupeaux, des herbages et des étables des claires, des houlettes, du soc, dont la poésie grecque et latine a transmis l'image et le chant, et que les salons du sont à présent oublié ».

<sup>10</sup> Richard, Jean Baptiste-Claude, Abate di Saint-Non. *Voyage pittoresque ou description des royaumes de Naples et de Sicile*. Vol. 3, Paris: Lafosse, 1783, 43-45: « Nous ne pouvions quitter un lieu dont la fraîcheur et la singularité avoient autant de charmes pour nous, et nous nous mîmes tous à dessiner et à en prendre plusieurs Vues, sous différents aspects en dedans et en dehors. Mais on doit sentir qu'un effet qui tient en plus grande partie à la magie de la couleur, ne peut être rendu que très-imparfaitement dans des Dessins, et surtout par des Gravures qui ne sont point coloriées ».

<sup>11</sup> Il riferimento a Caen di André Maurel può essere relativo alla chiesa di Santo Stefano, significativo esempio dell'architettura romanica.

riscuote Monte Sant'Angelo meta di pellegrinaggi di re ed imperatori come ricorda Claude Jordan<sup>12</sup>. E che dire della Puglia così presente per le vicende dell'impero romano? A Brindisi Cesare assediò Pompeo, a Taranto si ebbe l'assedio di Annibale, a Canne della Battaglia, presso quello che è definito il campo del sangue, la sconfitta romana dinanzi all'esercito cartaginese guidato da Annibale. L'immaginario "storico" della Puglia spazia così dalle vicende romane alle vicende dei normanni, da Federico II a Carlo d'Angiò, da Manfredi sino all'invasione Turca di Otranto. Ma il tacco d'Italia è considerato ormai dai viaggiatori francesi – al netto delle battaglie, degli assedi e delle invasioni – una terra d'esotismo, con venature romantiche suggerite dalla solitudine della punta estrema della Puglia così come scrive ancora Charles Yriarte: «assis sur les roches, pendant de vue la région cultivée où s'élèvent les villas entre Leuca et Otrante, ou se seul abandonné dans cette solitude, seulement troublée par le doux murmure du flot de la mer Ionienne qui vient mêler ses eaux à celles de l'Adriatique»<sup>13</sup>. Siamo a Leuca: come riporta Henry Castela, lì dove risiede ora il santuario Santa Maria De Finibus Terrae vi era in passato il tempio dell'impudica Venere, secondo la leggenda terra di approdo di Enea<sup>14</sup>. E così il promontorio pugliese che abbraccia i due mari, quello adriatico e quello ionico, significa così l'abbraccio di due storie, quella di Troia e quella di Roma e tutto si fonde in un richiamo lontano nel tempo fra storia e mito.

### 3. OLTRE LE COLONNE D'ERCOLE DI NAPOLI: LA VALORIZZAZIONE DELLA PUGLIA NEL SECONDO NOVECENTO

Questo patrimonio letterario fu apprezzato e conosciuto in modo sempre più crescente a partire dal secondo dopoguerra. Era ormai chiaro che a fronte della minore importanza attribuita al Sud Italia occorreva scalfire un pregiudizio: non era forse il Sud Italia altrettanto affascinante quanto altri territori più celebrati e visitati? Solo il viaggio al Sud, l'esperienza visiva e concreta, può sedurre i più scettici: tale è la tesi di Georges et Huguette Vallet che firmano uno tra i primi volumi della seconda metà del Novecento (1966) teso a promuovere il viaggio a sud oltre Napoli, in Lucania, Puglia e Calabria<sup>15</sup>. L'immagine ricorrente nei testi del Novecento inerenti la Puglia è tratta dalla mitologia: si inizia a scrivere e a dibattere della Puglia a partire dal rovesciamento della credenza diffusa che l'Europa finisse a Napoli e che nel capoluogo partenopeo avessero preso dimora le imponenti colonne d'Ercole a monito per i viaggiatori del confine ultimo del continente. Superate le colonne il viaggiatore assumeva su di sé rischi e pericoli, disagi e disavventure. Fu Augustin François Creuzé de Lesser a inaugurare i testi di viaggio del diciannovesimo secolo evidenziando come «L'Europe finit à Naples et même elle y finit assez mal: la Calabre, la Sicile, tout le reste est de l'Afrique.»<sup>16</sup>. In virtù di questo pregiudizio, tutte le pubblicazioni del secondo Novecento afferenti all'editoria turistica restituiscono – a ragion veduta – forti dosi di stupore nel constatare la bellezza del Sud Italia, il suo legame profondo con la storia, il suo innato valore paesaggistico.

Per Maria Brandon-Albini – curatrice nel 1971 di una delle prime guide turistiche francesi dedicate al Grande Sud Italia, si tratta di «un retour aux sources d'une civilisation universelle et plurimillénaire. Ces racines humaines archaïque, voire intemporelles, comme dit Carlo Levi, créent des âmes, en apparence farouches, qui bientôt s'ouvrent devant nous en nous laissant deviner leurs tourments, leur délicatesse, leur dignité presque malade, leur hospitalité amicale digne de l'époque d'Homère. Au fait, ne sommes-nous pas dans la Grande-Grèce?»<sup>17</sup>.

Singolare la testimonianza di Marie-Joséphine Sellier che, nel suo *L'Italie inconnue d'Eboli à Termoli* edito nel 1988, racconta dei suoi primi viaggi in Puglia e Basilicata seppur sconsigliata a più riprese per via della pericolosità e dei disagi che comportava un viaggio a sud di Napoli: «Non seulement je n'ai pas été mangée par les loups ni enlevée par les bandits, mais je peux dire que jamais je n'ai fait en Italie des voyages aussi intéressants, passionnantes, même, à tous points de vue»<sup>18</sup>.

Non solo testimonianze di viaggio e guide turistiche, nel secondo Novecento e oltre, la Puglia attira l'attenzione di diversi scrittori francesi. Il poeta Gabriel Faure nel suo *Viellard de Tarente*<sup>19</sup> riprende l'episodio del *senex* virgiliano delle *Georgiche* che lasciò la Cilicia per vivere nei pressi di Taranto e coltivare il proprio giardino. Questa evocazione virgiliana lo spinge a tornare in Italia e ad affacciarsi per la prima volta in Puglia e scrivere una «rêverie» su Taranto, sulla saggezza della vecchiaia, sulla felicità della vita agreste. André Pieyre de Mandiargues a fine anni Cinquanta descrive la Puglia come

<sup>12</sup> *Voyages historiques de l'Europe*. Tome III, Paris: N. Le Gras, 1693.

<sup>13</sup> Yriarte, Charles. *Les Bords de l'Adriatique et le Montenegro*. Venise. *L'Istrie. Le Quarnero. La Dalmatie. Le Montenegro et la rive italienne*. Paris: Hachette, 1878, 593-626, chap. XI.

<sup>14</sup> Castela, Henry. *Le Saint Voyage de Hierusalem et Mont Sinay*. Bordeaux-Paris: Sonnius, 1603.

<sup>15</sup> Vallet, Georges et Huguette. *Nous partons pour... Naples et l'Italie du Sud*. Paris: Presses Universitaires de France, 1966.

<sup>16</sup> Creuzé de Lesser, Augustin François. *Voyage en Italie et en Sicile, fait en 1801-et 1802*. Paris: Didot, 1806, 96.

<sup>17</sup> Brandon-Albini, Maria. *Le Grand Sud italien*. Paris: Hachette, 1971, 10.

<sup>18</sup> Sellier, Marie-Joséphine. *L'Italie inconnue d'Eboli à Termoli*. Paris: La pensée universelle, 1988, 10.

<sup>19</sup> Faure, Gabriel. *Le Viellard de Tarente suivi des Délices de Juin*. Paris: Les Horizons de France, 1956.



la terra «qui nous donne à si haut point l'agréable impression de nous être *éloignés* dans le temps et dans l'espace»<sup>20</sup>. L'idea di Puglia come luogo dove spazio e tempo si arrestano è in comune con un altro scrittore, il contemporaneo Patrick Dréhan. Nel suo romanzo *Le Voyage en Italie* (Paris, Editions Sydney Laurent, 2020) Otranto e le località Salento rappresentano le città « où s'arrête le temps, où commence le permis de rêver »<sup>21</sup>. Anche il Gargano è ben rappresentato in testi letterari entrambi vincitori del prestigioso premio Goncourt: nel 1957 Roger Vailland ambienta il suo *La Lois* (Paris, Gallimard, 1957) sul promontorio pugliese<sup>22</sup>, così come Peschici e Monte Sant'Angelo ispirano Montepuccio, luogo dove è ambientato il *Soleil des Scorta*<sup>23</sup> di Laurent Gaudé. Il romanzo *Belle infidèle*<sup>24</sup> di Romane Lafore del 2019 spazia, invece, tra Monopoli e il Salento, una storia d'amore che ha come sfondo il mare, il sole e la luce pugliese<sup>25</sup>.

#### 4. CONSIDERAZIONI FINALI SU DIGITAL HUMANITIES E TURISMO LETTERARIO

La promozione dell'immaginario francese della Puglia permette di fare alcune considerazioni sullo stato dell'arte circa gli studi sul Grand Tour e sui progetti scientifici afferenti alla sfera delle Digital Humanities. Il caso di studio dei testi dei viaggiatori nel Trentino Alto Adige<sup>26</sup> ha evidenziato, ancora una volta, la svolta impressa dagli strumenti digitali per lo studio delle analisi spaziali in un contesto di opere letterarie. Sebbene non vi sia ad oggi una metodologia universale da applicare al trattamento della "digital history" – i curatori del Progetto di Stanford "Mapping the Republic of Letters" non ritengono, ad esempio, di dover evidenziare a livello geografico i confini esatti con il GIS a vantaggio della definizione delle "zone culturali"<sup>27</sup> – unanime è la considerazione circa i vantaggi nel poter disporre di un corpus digitale di letteratura odeporea e di un database georeferenziato circa i luoghi descritti e citati dai viaggiatori.

Vi è, infine, una questione culturale molto profonda da approcciare, una questione suggerita, anzi provocata, da Franco Moretti nel *Falso Movimento*<sup>28</sup>. È vero che in gran parte gli studi di Digital Humanities hanno reciso il legame con le grandi teorie letterarie del Novecento e che a farne le spese sia stato il concetto di forma? È vero che se una teoria non sovrasta l'uso e gli strumenti di analisi dei dati, saranno i luoghi comuni a prendere il sopravvento? Probabilmente sì, i progetti di ricerca non devono essere solo guidati dai dati ma possono in qualche modo evitare di ignorare le grandi teorie letterarie e le questioni di forma. È per questo motivo che questo progetto non ignora alcune considerazioni – in limine – sullo status del canone della letteratura di viaggio, delle forme, delle costanti e delle varianti dei testi del *Grand Tour*, all'interno del quale prende forma questo *Petit Tour* a tratti in forme originali e a tratti in forme "prese a prestito".

Sebbene studiosi come Todd Presner e David Shepard immaginino il futuro di questi studi "spaziali" di turismo letterario-geografico nelle mappe immersive avatar-based, nonché nel mondo virtuale del gaming<sup>29</sup>, ritengo, in conclusione, che ad oggi queste iniziative raggiungano già in modo ottimale il loro scopo soprattutto se vi è un legame forte con il territorio e con le istituzioni che ne governano i processi di sviluppo integrato. Le potenzialità dell'esito del progetto "L'immaginario francese della Puglia" non risiedono tanto nella possibilità della lettura di testi digitalizzati – la "memoria" digitale – quanto nella promozione digitale di una Puglia già apprezzata dai francesi e descritta secondo categorie dell'immaginario che ben si prestano a suggerimenti di itinerari originali rispetto alle proposte tradizionali delle guide turistiche, dunque nell'accesso

<sup>20</sup> Pieyre de Mandiargues, André. *Le Belvédère*. Paris : Bernard Grasset, 1958, 112. Nello stesso anno Mandiargues dedicherà un breve capitolo a Castel del Monte – *L'Espion des Pouilles* – nel suo *Cadran Lunaire*. Paris : Robert Laffont, 133-138.

<sup>21</sup> Dréhan, Patrick. *Le Voyage en Italie*. Paris: Editions Sydney Laurent, 2020, 78.

<sup>22</sup> Porto Manacore del romanzo di Vailland non è altro che Peschici.

<sup>23</sup> Gaudé, Laurent. *Le soleil des Scorta*. Arles: Actes Sud, 2004.

<sup>24</sup> Lafore, Romane. *Belle infidèle*. Paris: Stock, le Livre de Poche, 2019.

<sup>25</sup> Una parentesi a parte, infine, è costituita da testi di autori francesi che narrano la storia delle loro famiglie d'origine. Giovanni-Michel Del Franco nel 2000 (*Accadia, vie et mort d'un village du sud de l'Italie*. Argenteuil: Le Chant des Hommes) racconta il suo viaggio sui passi del padre originario di Candela ed emigrato per motivi economici; Hervé Deguine nel 2011 (« *Avanti !* ». *Histoire d'une famille italienne venue des Pouilles 1780-1947*. Paris : Bonaventure) narra dei suoi nonni italiani e della vita ad Andria tra guerra e dopoguerra. Entrambi i testi evidenziano l'aspetto rurale dei paesi interni pugliesi, una civiltà quasi arcaica improntata su rapporti di convivenza fatti di pura semplicità.

<sup>26</sup> Cfr. Dai Prà, Elena, e Gabellieri, Nicola. "Mapping the Grand Tour Travel Writings: a GIS-Based Inventorying and Spatial Analysis for Digital Humanities in Trentino-Alto Adige, Italy (XVI-XIX c.)". *Literary Geographies*, 7, 2, 2021, 251-274.

<sup>27</sup> Cfr. Edelstein, Dan, Findlen Paula, Ceserani Giovanna, Winterer Caroline, e Coleman Nicole. "Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project". *American Historical Review*, 122, 2, April 2017, 421 : «At the geographical level, we do not place much importance on exact boundaries or even precise locations; for this reason, we have not used GIS (geographic information systems) in our visualizations, and we tend to prefer abstract maps over historical ones. We take as a given that cultural zones have fuzzy borders and can overlap: eighteenth- century Geneva, for instance, could be seen as belonging to both a Swiss Protestant and a French cultural zone. We use the term "cultural zone" rather than "cultural space" in order to stress its imprecise, often amorphous, edges. A zone also covers a large area, which is why we chose this term over "place." It can indeed capture connections between places in a given period.».

<sup>28</sup> Moretti, Franco. *Falso movimento. La svolta quantitativa nello studio della letteratura*. Milano: Nottetempo, 2022.

<sup>29</sup> Presner, Todd, e Shepard, David. "Mapping the Geospatial Turn", in Susan Schreibman, Ray Siemens, John Unsworth (eds.), *A new Companion to Digital Humanities*. Blackwell: Wiley, 2016, 201-212.

digitale in modo semplice ed intuitivo ad un patrimonio letterario notevole e di sicuro impatto sul turismo esperienziale. Tutto ciò, inoltre, trova ampio riscontro nell’iniziativa regionale “Puglia 365, Piano Strategico Turismo Cultura al 2030” che prevede tra gli obiettivi futuri per la Puglia: l’implementazione della “brand reputation” per diversificazione, promuovendo arte, cultura e tradizioni; la delocalizzazione delle proposte turistiche, prevedendo itinerari distanti dai soliti circuiti che prediligono il “prodotto mare”<sup>30</sup>.

Un patrimonio culturale immateriale diventa quindi tangibile nel momento in cui le istituzioni – come Puglia Promozione, agenzia della Regione Puglia – sposano le finalità della ricerca mettendo in connessione il mondo accademico con gli operatori del settore, in questo caso operatori turistici.

In tal senso il progetto “L’immaginario francese della Puglia” si configura come una “deep map” secondo la definizione di David J. Bodenhamer<sup>31</sup> in quanto le mappe dei viaggiatori e scrittori francesi non si limitano a disegnare un patrimonio tangibile e materiale, ma includono dimensioni discorsive e ideologiche dei luoghi raccontati come le sfumature di approccio raccolte secondo le categorie dell’immaginario. La deep map della Puglia è quindi – come osserva lo stesso Bodenhamer – una piattaforma, un processo e un prodotto. In quanto piattaforma questa deep map relaziona il dato letterario con spazio e tempo; in quanto processo traccia una nuova ed originale narrazione per i turisti francesi; in quanto prodotto è una rappresentazione visiva di questa nostra indagine che si presta ad una discussione più ampia sul modo di promuovere la Puglia all’estero.

## BIBLIOGRAFIA

- [1] Cooper David., Donaldson Christopher, e Murrieta-Flores Patricia. *Literary mapping in the digital age*. London: Routledge, 2016.
- [2] Dai Prà, Elena, e Gabellieri, Nicola. “Mapping the Grand Tour Travel Writings: a GIS-Based Inventorying and Spatial Analysis for Digital Humanities in Trentino-Alto Adige, Italy (XVI-XIX c.)”. *Literary Geographies*, 7, 2, (2021): 251-274.
- [3] Dotoli Giovanni, e Fiorino Fulvia. *Viaggiatori francesi in Puglia*. 8 voll. Fasano: Schena, 1985-1999.
- [4] Edelstein, Dan, Findlen Paula, Ceserani Giovanna, Winterer Caroline, e Coleman Nicole. “Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project”. *American Historical Review*, 122, 2, (April 2017).
- [5] Engberg-Pedersen, Anders (ed.). *Literature and cartography: Theories, histories, genres*. MIT Press, 2017.
- [6] Gizard, Christopher. *Conception d’une méthode hybride d’extraction d’informations géographique à partir de donne textuelles*. Université Claude Bernard Lyon 1, LabEx IMU.
- [7] Hermann, Maria Luisa Semeraro, Semeraro Angelo, e Semeraro Raffaele. *Viaggiatori in Puglia dalle origini alla fine dell’Ottocento*. Brindisi: Schena, 2000.
- [8] Menichelli, Giancarlo. *Viaggiatori francesi reali o immaginari nell’Italia dell’Ottocento*. Roma: Edizioni di Storia e Letteratura, 2012.
- [9] Moretti, Franco. *Falso movimento. La svolta quantitativa nello studio della letteratura*. Milano: Nottetempo, 2022.
- [10] Moretti, Franco. *La letteratura vista da lontano*. Torino: Einaudi, 2005.
- [11] Piatti, Barbara. “Literary Cartography: Mapping as Methods”, in Engberg-Pedersen Anders (ed.), *Literature and cartography: Theories, histories, genres*. 45-72, MIT Press, 2017.
- [12] Presner, Todd, e Shepard David. “Mapping the Geospatial Turn”, in Susan Schreibman, Ray Siemens, John Unsworth (eds.), *A new Companion to Digital Humanities*. 201-212, Blackwell: Wiley, 2016.
- [13] Semeraro, Angelo. *Viaggiatori in Puglia dalle origini fino alla fine dell’Ottocento*. Fasano: Schena, 1991.
- [14] Semeraro, Raffaele. *Viaggiatori in Puglia*. Fasano: Schena, 1991.

---

<sup>30</sup> *Puglia 365, Piano Strategico Turismo Cultura al 2030*. Position paper “Arte e Cultura”, <https://www.puglia365.it>

<sup>31</sup> Bodenhamer, David J. “Making the Invisible Visible Place, Spatial Stories and Deep Maps”, in Cooper David., Donaldson Christopher, e Murrieta-Flores Patricia (eds.), *Literary Mapping in the Digital Age*. London: Routledge, 2016, 212-213.

## **5. Workshop**

# Workshop sulle edizioni digitali: preparazione con codifica XML TEI e visualizzazione con il software EVT

Giulia D'Agostino<sup>1</sup>, Giulia Fabbris<sup>2</sup>, Roberto Rosselli Del Turco<sup>3</sup>

<sup>1</sup> Università degli Studi di Verona, Italia – giulia.dagostino@univr.it

<sup>2</sup> Università Ca' Foscari Venezia, Italia – giulia.fabbris@unive.it

<sup>3</sup> Università di Torino, Italia – roberto.rossellidelturco@unito.it

## ABSTRACT

Per il dodicesimo convegno annuale AIUCD “La memoria digitale. Forme del testo e organizzazione della conoscenza” proponiamo un workshop sulla preparazione di edizioni digitali XML TEI e la loro visualizzazione con il software EVT. Gli obiettivi principali del workshop sono due. In primo luogo, vorremmo stimolare riflessioni su trattamento e rappresentazione di dati testuali. In secondo luogo, miriamo a mostrare nel concreto come si costruisce, passo per passo, un'edizione digitale e come può essere visualizzata. Per quest'ultimo fine, insegneremo ad utilizzare EVT in modo che gli iscritti possano poi lavorare autonomamente e ottenere più consapevolezza del rapporto tra dati di partenza e software. Il workshop è rivolto a tutti coloro che sono interessati ad un esempio di *workflow* di gestione del materiale testuale.

## PAROLE CHIAVE

Filologia digitale; modellazione di dati; software di visualizzazione; XML TEI

## 1. INTRODUZIONE

Questo workshop costituisce un'introduzione ai temi della filologia digitale e delle edizioni digitali, con particolare attenzione a problemi metodologici che hanno a che fare con organizzazione e visualizzazione di dati testuali. Include un approccio pratico, con una sessione dedicata alla preparazione e visualizzazione di un'edizione critica. Al termine del corso i partecipanti avranno conseguito una formazione di base per la creazione e pubblicazione di edizioni digitali con EVT sulla base di una codifica secondo lo standard XML TEI.

Gli iscritti potranno scaricare un archivio contenente il materiale necessario per seguire le lezioni e portare a termine l'esercitazione. In alternativa, incoraggiamo chi avesse un proprio progetto di edizione a portare il proprio materiale di base: immagini, trascrizione in formato testo e, se disponibile, il testo già codificato in XML TEI. Non sono richieste conoscenze particolari, ma è certamente utile una competenza anche minima nell'uso di editor di testo, in particolare editor XML (v. sotto).

Il software utilizzato per la visualizzazione è **EVT – Edition Visualization Technology** (<http://evt.labcd.unipi.it/>), in particolare la versione 2 creata specificamente per edizioni critiche e diplomatiche. Dato che EVT accetta in input documenti nel formato XML TEI, buona parte del corso sarà dedicata proprio alle nozioni fondamentali di questo linguaggio in relazione alla preparazione di edizioni critiche, concentrandoci in particolare sulla marcatura dell'apparato critico e riflettendo su vantaggi e problemi del markup.

L'obiettivo principale della TEI è, infatti, fornire delle linee guida per l'annotazione e la strutturazione di dati testuali in formato digitale. Per questo motivo vorremmo che la riflessione sulla codifica fosse il punto di partenza del workshop, per mostrare come l'interpretazione del testo con l'obiettivo di restituirlo ai fruitori sia il principio alla base del processo editoriale. L'introduzione teorica includerà non solo questa parte, ma anche una discussione su come presentare l'edizione al pubblico: il supporto con cui viene visualizzata l'edizione e l'interfaccia utente sono a tutti gli effetti parte del lavoro interpretativo e influenzano la ricezione del testo edito, oltre a determinare se un potenziale utente si soffermerà ad utilizzare l'edizione digitale oppure no.



Figura 1. EVT 2 beta2 nella view collazione.

Dato tutto ciò, puntiamo a mostrare l'utilità di EVT, che permette di visualizzare le edizioni XML TEI in un'applicazione web accessibile dal browser: pensato come *tool* generico, basato su standard web diffusi (HTML 5, CSS, JavaScript), e *open source*, EVT è pronto all'uso subito dopo l'installazione, e richiede pochi passaggi per pubblicare un'edizione digitale. Questo strumento, infatti, è stato progettato per essere di facile utilizzo non soltanto per l'utente finale, ma anche per l'utente editor, in modo da permettere l'accesso a una pubblicazione sul web anche a giovani ricercatori e/o studiosi che non dispongono dei fondi necessari per soluzioni più costose (anche per quanto riguarda la manutenzione nel lungo periodo). Durante la parte pratica vedremo in quale misura lo studioso deve/può intervenire per configurare e personalizzare EVT, invitando i partecipanti a testare le sue numerose funzionalità in base agli esercizi di codifica svolti nella prima parte del workshop o a materiali sui quali stanno lavorando.

## 2. SOFTWARE UTILIZZATO

Per la frequenza del corso è necessario disporre sul proprio computer dei seguenti dati e programmi:

- un editor XML;
- la cartella con i materiali scaricabili dal link comunicato all'inizio del workshop;
- la versione beta 2 del software EVT 2 (se non inclusa nella cartella materiali).

L'editor XML più diffuso per la marcatura dei testi è **Oxygen XML Editor**: lo si può scaricare e registrarsi in maniera da avere un periodo di prova gratuito di trenta giorni (<http://oxygenxml.com/download.html>). Un'alternativa gratuita e *open source*, anche se con qualche limitazione, è **XML Copy Editor** (<http://xml-copy-editor.sourceforge.net/>). Al momento attuale la soluzione open source migliore è l'editor **Visual Studio Code** (<https://code.visualstudio.com/>) aggiungendo l'estensione *Scholarly XML* creata da R. Viglianti: grazie a questa estensione, alla quale ne possono essere affiancate altre per la gestione di documenti XML, è possibile effettuare la validazione automatica basata su schemi del tipo RelaxNG. Nella cartella dei **materiali** (link comunicato all'inizio del workshop), che consigliamo di scaricare integralmente, troverete le slide impiegate durante il corso e un archivio `materiali.zip` che contiene:

- **doc**: documentazione riguardo lo schema di codifica utilizzato, un sunto (*cheatsheet*) della codifica TEI per l'apparato critico, un manuale per la codifica di testi medievali a cura di M. Burghart, un breve tutorial per usare EVT (`quick-start.pdf`);
- **edizione critica**: i testi che useremo per preparare una edizione critica;
- **edizione diplomatica**: i testi che useremo per preparare una edizione diplomatica;
- **esercizi**: eventuali esercizi proposti durante il corso;
- **evt**: la versione di EVT da usare per il corso (EVT 2 beta2);

- **xml**: documenti di partenza, e relativi schemi TEI, per fare la codifica dell'edizione.

Per installare qualsiasi versione di EVT sul proprio computer è sufficiente scompattare l'archivio scaricato in una cartella sul desktop o altrove sul proprio computer.

Per visualizzare edizioni prodotte con EVT in locale, sul proprio computer, si consiglia di usare la versione **60** di **Firefox ESR** (<https://ftp.mozilla.org/pub/firefox/releases/60.0esr/>) oppure **Chrome** lanciandolo **da riga di comando** con il parametro `--allow-file-access-from-files`. Per maggiori informazioni si veda sotto la nota riguardo il navigatore da usare in locale.

### 3. ISTRUZIONI DI BASE

Un breve sunto delle differenti fasi necessarie per creare un'edizione digitale usando EVT 2:

1. **codifica**: il primo passo per la preparazione della vostra edizione digitale è la codifica del testo per mezzo degli schemi XML TEI in modo da avere un documento TEI valido (si noti che grazie a XInclude si può avere un file principale che raccoglie molti altri documenti TEI separati);
2. **inserimento dati**: le differenti componenti dell'edizione devono essere successivamente copiate nelle relative cartelle all'interno della directory data, ad esempio le immagini a facciata singola devono andare in `data/images/single`;
3. **configurazione**: EVT deve poi essere configurato editando il file `config.json` (esiste un front-end sperimentale disponibile come web-app, ma non include le ultime caratteristiche) nella cartella `config`;
  - 3.1 come passo opzionale potete aggiungere le vostre regole CSS aggiungendole al file `custom-style.css` nella stessa cartella;
4. **testing**: per verificare i risultati, anche man mano che si fa la codifica del testo, basta caricare il file `index.html` in un navigatore; CAVEAT: non è più possibile farlo usando browser moderni come Firefox o Chrome, sfortunatamente (v. la sezione sulla navigazione di un'edizione locale più sotto).

### 4. DOCUMENTAZIONE E LINK UTILI

Per la codifica nel formato XML TEI sono fondamentali, oltre a quelli di base, i moduli relativi alla trascrizione di fonti primarie, alla descrizione del manoscritto e all'apparato critico, ben documentati nelle *Guidelines* TEI (<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>):

<i>Manuscript Description</i>	<a href="http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html">http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html</a>
<i>Representation of Primary Sources</i>	<a href="http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html">http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html</a>
<i>Critical Apparatus</i>	<a href="http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html">http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html</a>

Ci concentreremo soprattutto sul modulo *Critical Apparatus*, metodo *parallel segmentation* come descritto nelle *Guidelines*. Altri moduli, ad esempio *Names dates people and places*, *Linking segmentation and alignment* etc., possono essere utili per scopi specifici e saranno menzionati quando necessario.

#### Edizioni digitali basate su EVT 1

Vercelli Book Digitale:	<a href="http://vbd.humnet.unipi.it/beta2/">http://vbd.humnet.unipi.it/beta2/</a> .
Codice Pelavicino Digitale:	<a href="http://pelavicino.labcd.unipi.it/">http://pelavicino.labcd.unipi.it/</a> .
Humarec Manuscript Viewer. Marciana Gr. Z 11 (379):	<a href="http://humarec-viewer.vital-it.ch/">http://humarec-viewer.vital-it.ch/</a> .

#### Edizioni digitali basate su EVT 2

Pietro da Eboli, <i>De rebus Siculis Carmen</i> :	<a href="http://web.unibas.it/bup/evt2/pde/">http://web.unibas.it/bup/evt2/pde/</a>
Il Rotulo di San Teobaldo:	<a href="https://www.visitmudi.it/rotulo-di-san-teobaldo/">https://www.visitmudi.it/rotulo-di-san-teobaldo/</a>
Progetto VaSto <sup>1</sup> :	<a href="https://dharc-org.github.io/progetto-vasto/">https://dharc-org.github.io/progetto-vasto/</a> <a href="https://dharc-org.github.io/progetto-vasto/evt2beta2/">https://dharc-org.github.io/progetto-vasto/evt2beta2/</a>

---

<sup>1</sup> Edizione digitale della Storia fiorentina di Benedetto Varchi (1503-1565)

## Gli schemi di codifica TEI (Text Encoding Initiative)

Home page TEI: <http://www.tei-c.org/>.

Le *Guidelines*: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.

Il sito TEI by Example: <http://teibyexample.org/>.

### Nota sul navigatore per i test in locale

A partire dalla versione 67 gli sviluppatori di Firefox hanno deciso di seguire la stessa cauta politica già adottata dagli sviluppatori di Chrome (e altri navigatori), ovvero impedire il caricamento di risorse locali da parte di programmi JavaScript eseguiti nel navigatore. Questo, se garantisce in termini generali una maggior sicurezza, ha purtroppo lo spiacevole effetto collaterale di impedire il caricamento in locale di edizioni basate su EVT (che è appunto implementato in JavaScript) o altri programmi simili. In altre parole, se prima della v. 67 era possibile caricare il file `index.html` direttamente in Firefox e navigare un'edizione basata su EVT sul proprio computer, il che è molto comodo per controllare il progresso del proprio lavoro, adesso questo non è più possibile (nemmeno usando altri browser come Chrome, Opera, Edge etc.). Per fortuna ci sono alcune soluzioni che si possono adottare per fare questi test in locale:

- lanciare Chrome da riga di comando con il parametro `--allow-file-access-from-files`, poi fare CTRL+O per aprire il file `index.html` nella cartella di EVT, o direttamente *drag&drop* di quest'ultimo sulla finestra di Chrome;
- su Windows digitare queste istruzioni da PowerShell o altro terminale
  - o `cd "C:\Program Files (x86)\Google\Chrome\Application\"`
  - o `.\chrome.exe --allow-file-access-from-files`
- su Mac l'istruzione corrispondente è
  - o `open /Applications/Google\ Chrome.app --args --allow-file-access-from-files`
- oppure
  - o `open -a "Google Chrome" --args --allow-file-access-from-files`
- se si usa Visual Studio Code, ma anche in caso contrario, un metodo molto efficace per visualizzare edizioni in locale è installare l'estensione Live server (Five server), per utilizzarla bastano due passi:
  - o selezionate la cartella EVT (quella con il file `index.html`) scegliendo dal menù File → Open folder
  - o fate clic sul pulsante Go Live nella barra di stato (in basso a destra)
- NB: per essere sicuri che si apra l'edizione con EVT non deve essere visualizzato nessun documento XML o HTML o altro (v. screenshot sopra, basta tenere aperta la scheda di informazioni di una estensione) altrimenti Code cercherà di aprire quello nel server locale.

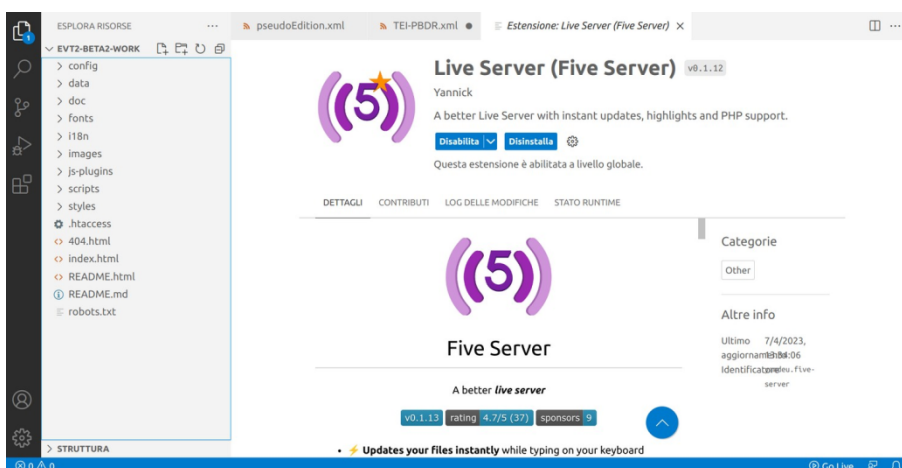


Figura 2.

- scaricare e installare **Firefox ESR** versione **60**: questa versione precede l'introduzione di questa nuova policy di sicurezza e si può installare in parallelo ad altre versioni di Firefox (NB: **non** è consigliato usare questa versione per le proprie esigenze normali di navigazione);



- installare su Firefox o Chrome un'estensione che funge da web server locale, ad esempio per Chrome è disponibile questa:
- <https://chrome.google.com/webstore/detail/web-server-for-chrome/ofhbbkphhbklhfoeikjpcbhmlcogigb/related>
- installare l'ambiente di sviluppo seguendo le istruzioni riportate sul repository GitHub (usare il branch develop, opzione consigliata solo a chi ha confidenza con git e lo sviluppo basato su GitHub): <https://github.com/evt-project/evt-viewer-angular>

La soluzione più semplice, soprattutto se avete già Chrome installato sul computer, è usare quest'ultimo lanciandolo da riga di comando. È anche possibile preparare uno script e/o un'icona dedicata a tale scopo sulla scrivania in modo da non dover ripetere il comando ogni volta:

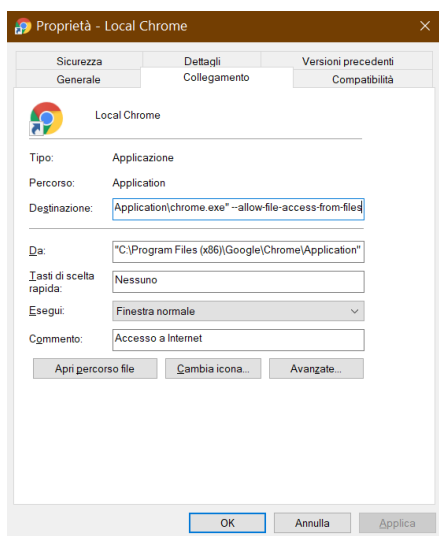


Figura 3. Windows.

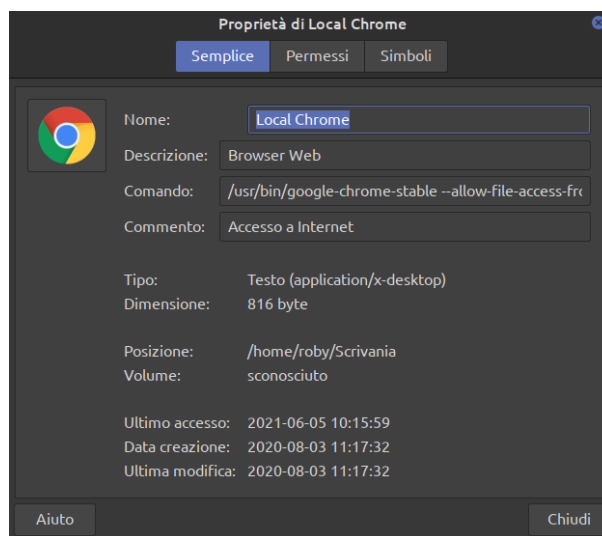


Figura 4. Linux.

In ogni caso il problema riguarda solo i test in locale: una volta caricata l'edizione finale su un qualsiasi server web non ci sono problemi ad accedere.

## BIBLIOGRAFIA

- [1] Burghart, Marjorie. "The TEI Critical Apparatus Toolbox: Empowering Textual Scholars through Display, Control, and Comparison Features." *Journal of the Text Encoding Initiative* [Online], no. 10 (2016). <http://jtei.revues.org/1520>; DOI: 10.4000/jtei.1520.
- [2] Burghart, Marjorie e Elena Pierazzo. Online course: Digital Scholarly Editions: Manuscripts, Texts, and TEI Encoding. Retrieved May 17, 2019, from <https://www.digitalmanuscripts.eu/digital-editing-of-medieval-texts-a-textbook/>. Include il manuale Digital Editing of Medieval Texts: A Textbook.
- [3] Cacioli, Giulia, Giacomo Cerretini, 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 - Digital cultures. Intersections: philosophy, arts, media. Proceedings of the 11th national conference*, Lecce, 1-3 June 2022, Fabio Ciraci, Giulia Miglietta e Carola Gatto (ed.). 212-217.2022. <http://amsacta.unibo.it/6848/>
- [4] Di Pietro, Chiara e Roberto Rosselli Del Turco. "Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions." In *Digital Scholarly Editions as Interfaces. Schriften des Instituts für Dokumentologie und Editorik — Band 12*, a cura di Roman Bleier, Martina Bürgermeister et al., 129-160. Norderstedt: BoD, 2018.
- [5] Driscoll, Matthew James e Elena Pierazzo, ed. *Digital Scholarly Editing: Theories and Practices*. Vol. 4. Digital Humanities Series. Open Book Publishers, 2016. <https://www.openbookpublishers.com/product/483>
- [6] O'Donnell, Daniel Paul. "Different Strokes, Same Folk: Designing the Multi-Form Digital Edition." *Literature Compass* 7.2 (2010): 110–119.
- [7] Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham, Surrey; Burlington, VT: Ashgate Publishing, Limited, 2015. <https://www.openbookpublishers.com/product/483>
- [8] Price, Kenneth M. "Electronic Scholarly Editions." In *A Companion to Digital Literary Studies*, 434-450. John Wiley & Sons, Ltd, 2013.

- [9] Robinson, Peter. "Towards a Theory of Digital Editions." *The Journal of the European Society for Textual Scholarship* (2013): 105–131.
- [10] Rosselli Del Turco, Roberto. "Il Vercelli Book Digitale: teoria e pratica di un progetto di edizione elettronica." In *Schede Umanistiche XXVIII* (2016): 181-200.
- [11] Rosselli Del Turco, Roberto. "Designing an advanced software tool for Digital Scholarly Editions. The inception and development of EVT (Edition Visualization Technology)." *Textual Cultures* 12.2 (2019): 91-111. <http://textual-cultures.org/wp-content/uploads/2018/04/document-8.pdf>
- [12] Siemens, Ray, Cara Leitch, Alex Garnett, Corina Koolen e Meagan Timney. "Toward Modeling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media." *Literary and Linguistic Computing* 27.4 (2012): 445–461.
- [13] TEI Consortium, eds. 2023. TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.6.0. Last updated on 23rd April 2023. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>

# Trascrivere, analizzare e pubblicare documenti d'archivio: Transkribus e l'intelligenza artificiale al servizio dello storico

Sara Mansutti<sup>1</sup>, Salvatore Spina<sup>2</sup>

<sup>1</sup> READ-COOP SCE, Austria – s.mansutti@readcoop.eu

<sup>2</sup> Università degli Studi di Catania, Italia – salvatore.spina@unict.it

## ABSTRACT

Il workshop ha l'obiettivo di riflettere sulle opportunità che l'Intelligenza Artificiale offre al ricercatore per lo studio di testi manoscritti e a stampa. Negli ultimi anni, l'uso dell'Intelligenza Artificiale si è fatto strada nell'ambito della ricerca umanistica e dei Beni Culturali, e uno dei maggiori campi di applicazione è stato quello della trascrizione automatica dei manoscritti. Utilizzando la piattaforma Transkribus – sviluppata nell'ambito del progetto europeo Horizon 2020 "READ" e oggi mantenuta e implementata dalla società cooperativa europea READ-COOP SCE –, i partecipanti impareranno ad allenare un modello di riconoscimento automatico del testo manoscritto (Handwritten Text Recognition), processare automaticamente layout complessi (es. tabelle e registri), arricchire manualmente o automaticamente il layout e il testo dei documenti tramite tag. Verranno, inoltre, mostrate varie possibilità per esportare e pubblicare online le trascrizioni, e per analizzarle al di fuori di Transkribus.

## PAROLE CHIAVE

Intelligenza Artificiale, Handwritten Text Recognition, Layout Analysis, Transkribus

## 1. OBIETTIVI DEL WORKSHOP

L'obiettivo principale del workshop è mostrare le potenzialità d'applicazione dell'Intelligenza Artificiale nella ricerca storica e archivistica, in particolare per la trascrizione di documenti manoscritti o stampe antiche. Il workshop inizierà con un'introduzione teorica sull'impatto che l'Intelligenza Artificiale potrebbe avere per lo storico e quali nuove possibilità di ricerca potrebbe aprire. Seguirà poi una parte più pratica in cui, avvalendosi della piattaforma Transkribus, verrà mostrato passo passo come addestrare e applicare modelli di Handwritten Text Recognition e analisi del layout. Infine, verranno presentate diverse possibilità per esportare le trascrizioni generate automaticamente da Transkribus, pubblicarle o analizzarle con altri strumenti. Al termine del workshop, i partecipanti avranno familiarizzato con la piattaforma Transkribus, acquisito gli strumenti per allenare un modello di IA e conosciuto diverse opzioni per pubblicare e analizzare le trascrizioni. Avranno inoltre acquisito una visione più ampia delle possibilità dell'IA in campo storico.

## 2. CONTENUTO DEL WORKSHOP

Dati, Big Data, Textual Analysis, Web Semantico, sono solo alcuni dei concetti fondanti del *digital turn*. Tutto diventa processabile, e la maggior parte degli oggetti si 'virtualizzano', portando l'esperienza di essi verso significati non più fisici. Le parole diventano 'enti'-dati equiparabili alle «astrazioni di second'ordine»<sup>1</sup> (i numeri), e come tali possono essere analizzate, attraverso algoritmi in grado di ampliare la nostra conoscenza. Ma se ciò è vero per edizioni testuali a stampa (convertibile in testo digitale leggibile dalla macchina grazie all'OCR) e per i testi "nativi digitali", il discorso cambia quando ci riferiamo all'enorme patrimonio archivistico, costituito da manoscritti di diverse epoche e da variegate strutture linguistiche e di linguaggio, su cui è impossibile effettuare analisi computazionali. Così, già dagli anni Cinquanta, emerge la necessità di sviluppare tecnologie in grado di "leggere" tali documenti.<sup>2</sup> Negli ultimi anni, sono stati finanziati vari progetti per lo sviluppo di software e piattaforme che rispondessero proprio a queste necessità: tra essi, possiamo

---

<sup>1</sup> Cartesio, *Meditazioni metafisiche*; Cartesio, *Discorso sul metodo*.

<sup>2</sup> Dimond, «Devices for Reading Handwritten Characters»; Dunley, «The National Archives - Machines Reading the Archive»; Fischer et al., «Automatic Transcription of Handwritten Medieval Documents»; Sivasankari e Victor, «Handwritten Text Recognition»; Seaward e Kallio, «Transkribus».

menzionare [tranScriptorium](#),<sup>3</sup> [Transkribus](#)<sup>4</sup> e [eScriptorium](#).<sup>5,6</sup> Grazie ad essi, l'IA riesce, infatti, a rendere i documenti archivistici dei testi digitali accessibili, sia all'utente del web, che all'analista-umanista che vuole operare su di essi. Parlare, oggi, di digitalizzazione delle carte archivistiche non può dunque prescindere dal "trasformarle in testi *machine-readable*"; ossia superare la semplice raccolta fotografica, che, seppur ha fatto sì che numerosissime fonti d'archivio venissero immesse nella rete Internet, ostacola un approccio metodologico fondato sull'uso di tecnologie informatiche.

Sono già numerosi gli archivi che utilizzano l'Handwritten Text Recognition (HTR) per rendere il loro patrimonio cartaceo digitalizzato maggiormente accessibile sia ai ricercatori e sia al pubblico generale, ampliando così le possibilità di studiare e scoprire il contenuto dei documenti manoscritti. Un esempio è offerto dal Gemeente Amsterdam Stadsarchief (Archivio della città di Amsterdam)<sup>7</sup> che dal 2016 sta lavorando al progetto *Alle Amsterdamse Akten* per la digitalizzazione, trascrizione e indicizzazione di tutti gli archivi notarili storici della città. Nel 2021, benché solo 1.5% dell'intera collezione di 20 milioni di immagini fosse stato trascritto tramite IA, vennero alla luce due menzioni di pagamenti fatti a Rembrandt per un ritratto da lui dipinto, fino ad allora sconosciuto.<sup>8</sup> Sebbene fin dal 2018 l'HTR sia stato recepito come un potenziale rivoluzionario nel ridefinire l'approccio dei ricercatori alle collezioni archivistiche,<sup>9</sup> non esistono ancora studi che ne abbiano misurato l'impatto. Un sondaggio fatto tra gli utenti di Transkribus, nel marzo-aprile 2019, ha evidenziato come le possibilità offerte dalla piattaforma – e dall'Handwritten Text Recognition in generale – abbiano incentivato la digitalizzazione e orientato la scelta dei materiali da digitalizzare verso documenti manoscritti piuttosto che fotografie o mappe. Il 33% dei partecipanti al sondaggio, inoltre, ha affermato che, senza l'accesso all'HTR, le trascrizioni dei documenti non sarebbero mai state effettuate, mentre il 40% ha sostenuto che le trascrizioni sarebbero state portate a termine ma in un tempo molto maggiore.<sup>10</sup>

Presentate dunque le possibilità che l'IA offre sia allo storico sia agli archivi, la parte pratica del workshop si focalizzerà sulla piattaforma Transkribus, sviluppata nell'ambito del progetto europeo Horizon 2020 "READ" e oggi mantenuta e implementata dalla società cooperativa europea READ-COOP SCE. L'applicazione web Transkribus non richiede alcuna installazione, rendendola quindi accessibile sia agli storici che sono interessati a processare i loro documenti, senza dover imparare nuove competenze informatiche, sia agli studenti e a coloro che si avvicinano per la prima volta alle Digital Humanities. Il primo aspetto fondamentale da trattare sarà addestramento di modelli per la trascrizione automatica dei documenti e l'analisi del layout.

Questa è la fase più importante nell'utilizzo dell'IA, quella in cui il ricercatore deve impegnare tempo ed energie, per poi poter processare efficacemente le collezioni che vuole studiare ed ottenere risultati soddisfacenti. Uno dei maggiori vantaggi della piattaforma Transkribus è proprio la possibilità di addestrare modelli di IA specifici per i propri documenti, senza alcuna competenza richiesta in linguaggi di programmazioni o machine-learning. I modelli HTR sono quelli più conosciuti, ma è anche possibile allenarne altri per riconoscere automaticamente layout complessi (es. cartoline, moduli, mappe, annotazioni marginali...), per estrarre dati da tabelle e registri, e per taggare automaticamente sia elementi del layout, sia parole del testo – ad esempio, le abbreviazioni. I partecipanti impareranno a conoscere tutte queste diverse tipologie di modelli addestrabili all'interno di Transkribus e a comprendere come combinarle per processare i loro documenti. A partire da una serie di esempi, verrà mostrato come preparare e annotare le pagine di Ground Truth, come fare il training, quali risultati aspettarsi e come valutare l'efficacia di un modello.

Una volta terminato il training, i modelli possono essere applicati per trascrivere nuove pagine dalle caratteristiche simili a quelle usate durante l'addestramento. Le trascrizioni così generate – eventualmente, ma non necessariamente corrette manualmente –, possono poi essere utilizzate per diversi scopi: il principale è l'esportazione dei testi in formati digitali che possano garantire, da un lato, la diffusione – e quindi l'accesso e la consultazione – attraverso il Web; da un altro lato, l'analisi testuale attraverso strumenti quali Voyant Tool, Recogito, e altri ancora che si fondano sull'uso di particolari formati, come il semplice ".txt", o i più complessi "TEI-XML". Transkribus consente, inoltre, di esportare le collezioni in

---

<sup>3</sup> Sánchez et al., 'Handwritten Text Recognition for Historical Documents in the Transcriptorium Project'.

<sup>4</sup> Kahle et al., «Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents»; Muehlberger et al., «Transforming Scholarship in the Archives through Handwritten Text Recognition».

<sup>5</sup> Kiessling et al., 'eScriptorium: An Open Source Platform for Historical Document Analysis'.

<sup>6</sup> Vari articoli hanno analizzato le differenze tra i software e sistemi di riconoscimento HTR; in particolare i due qui citati mettono a confronto Transkribus ed eScriptorium: Huff and Stöbener, 'Projekt OCR-BW'; Maarand et al., 'A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian'. Il seguente resoconto della giornata di studio tenutasi presso la Bibliothèque nationale de France il 9 maggio 2022 offre invece una visione su come le due piattaforme si stanno sviluppando: Gautier et al., 'Compte-rendu de la journée d'étude « Point HTR 2022 » Transkribus / eScriptorium'.

<sup>7</sup> <https://transkribus.eu/r/amsterdam-city-archives/#/>

<sup>8</sup> Ponte e Schmitz, «Rembrandt Paints Master Carpenter Jacob Wesselsz Wiltingh.»

<sup>9</sup> Dunley, «The National Archives - Machines Reading the Archive».

<sup>10</sup> Terras, «Inviting AI into the Archives».

PDF ricercabili che, da un lato, garantiscono un approccio “analogico” alla fonte, e, dall’altro, consentono una prima dinamicità dello studio di essa, ossia la ricerca nel corpo del testo. I PDF di Transkribus possono, successivamente, diventare dei “Flipbook” consultabili attraverso portali e siti Web (es. <https://www.biscariepistolography.it>), rispondendo, in tal modo, ad uno dei principi cardine dell’era digitale: la diffusione capillare dei prodotti della ricerca. Tra le varie opzioni, è possibile anche esportare le trascrizioni in formato TEI-XML. Lo standard TEI è uno strumento importante per la conservazione e l’interoperabilità delle edizioni digitali, nonché per facilitare la loro pubblicazione online tramite tools come EVT e TEI Publisher. Per chi volesse invece pubblicare immagini e trascrizioni direttamente da Transkribus, senza scaricarle e ricaricarle su un altro server, la piattaforma [Read&Search](#) facilita questo processo, permettendo di pubblicare le proprie collezioni online, rendendole accessibili a tutti o previa registrazione.

L’IA, come mostra Transkribus, rappresenta, quindi, una possibilità innovativa per l’Umanistica Digitale, con particolare riferimento alle progettualità archivistiche e storiche, oltre che filologiche e linguistiche. La trascrizione automatica garantirebbe, infatti, un accesso a fonti digitali sempre più numerose e, soprattutto, accelererebbe i lavori di digitalizzazione dei fondi archivistici. Dal suo canto, lo storico – così come il linguista ed il filologo – avrebbe la possibilità di lavorare su più fonti relative al suo progetto di ricerca. Ad esempio, per descrivere la pandemia di peste del 1630, gli storici hanno lavorato su una documentazione che rappresentava non più di tre giorni della vicenda. L’acquisizione digitale di una documentazione più corposa, durante il progetto “Venice Time Machine” (Venezia-Losanna),<sup>11</sup> ha consentito, invece, uno studio su tre interi anni dell’evento pestilenziale.<sup>12</sup>

Ottenute le trascrizioni, l’umanista può, infine, avvalersi di ulteriori strumenti per soddisfare le necessità della sua ricerca e analizzare grandi quantità di testi, andando oltre il *close reading* e l’interpretazione. Tra questi, sicuramente Keyphrase Digger<sup>13</sup> rappresenta un valido esempio: un tool in grado di estrapolare “frasi-chiave” da corpi di testi (quali potrebbero essere le lettere di un fondo “Corrispondenza” di una casata nobile), con la finalità di mostrare – in una rappresentazione statistico-linguistica – delle informazioni utili allo storico per approfondire aspetti della sua ricerca che sarebbero emersi, magari, dopo un lungo tempo d’analisi. Oppure, la nota – oggi – ChatGPT (OpenAI),<sup>14</sup> i cui modelli di training consentono a questa IA di correggere i testi esportati da Transkribus, in diverse lingue moderne, così che lo storico possa lavorare su documenti la cui struttura linguistica non presenta errori che impedirebbero ad altri tools di operare efficacemente.

### 3. STRUTTURA

Il workshop (due ore) è stato pensato secondo questa struttura:

1. Lo scoglio dei testi manoscritti e le opportunità dell’Intelligenza Artificiale per lo storico (20 minuti)
2. L’impatto dell’Intelligenza Artificiale negli archivi (20 minuti)
3. Come addestrare modelli di Intelligenza Artificiale con Transkribus (40 minuti)
4. Come esportare le trascrizioni e pubblicarle (20 minuti)
5. Come analizzare le trascrizioni al di fuori di Transkribus (10 minuti)
6. Domande (10 minuti)

Non è richiesta nessuna competenza o conoscenza pregressa per partecipare al workshop. Chiediamo ai partecipanti di venire con il proprio laptop e di registrarsi alla piattaforma [Transkribus](#).

### BIBLIOGRAFIA

- [1] Alshater, Muneer. «Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT». SSRN Electronic Journal, 2022. <https://doi.org/10.2139/ssrn.4312358>
- [2] Archeomatica, Redazione. «Venice Time Machine: quando i Big Data sposano la cultura». Archeomatica - Cultural Heritage Technology, 29 giugno 2017. <https://www.archeomatica.it/ict-beni-culturali/venice-time-machine-quando-i-big-data-sposano-la-cultura>
- [3] Biswas, Som. «ChatGPT and the Future of Medical Writing». Radiology, 2 febbraio 2023, 223312. <https://doi.org/10.1148/radiol.223312>
- [4] Cartesio, Renato. Discorso sul metodo. Armando Editore, 1999.
- [5] ———. Meditazioni metafisiche. Armando Editore, 2003.

<sup>11</sup> Kaplan, «The Venice Time Machine»; Archeomatica, «Venice Time Machine»; Kaplan e di Lenardo, «Big Data of the Past».

<sup>12</sup> Lazzari et al., «A Digital Reconstruction of the 1630–1631 Large Plague Outbreak in Venice».

<sup>13</sup> Moretti, Sprugnoli, e Tonelli, «Digging in the Dirt».

<sup>14</sup> Biswas, «ChatGPT and the Future of Medical Writing»; Jiao et al., «Is ChatGPT A Good Translator?»; Pavlik, «Collaborating With ChatGPT»; Sobania et al., «An Analysis of the Automatic Bug Fixing Performance of ChatGPT»; Zhai, «ChatGPT User Experience»; Alshater, «Exploring the Role of Artificial Intelligence in Enhancing Academic Performance».

- [6] Dimond, T. L. «Devices for Reading Handwritten Characters». In Papers and Discussions Presented at the December 9-13, 1957, Eastern Joint Computer Conference: Computers with Deadlines to Meet on XX - IRE-ACM-AIEE '57 (Eastern), 232–37. Washington, D.C.: ACM Press, 1958. <https://doi.org/10.1145/1457720.1457765>
- [7] Dunley, Richard. «Machines Reading the Archive: Handwritten Text Recognition Software». Text. The National Archives blog. The National Archives, 19 marzo 2018. <https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/>
- [8] Fischer, Andreas, Markus Wuthrich, Marcus Liwicki, Volkmar Frinken, Horst Bunke, Gabriel Viehhauser, e Michael Stolz. «Automatic Transcription of Handwritten Medieval Documents». In 2009 15th International Conference on Virtual Systems and Multimedia, 137–42. Vienna, Austria: IEEE, 2009. <https://doi.org/10.1109/VSM.2009.26>
- [9] Gautier, Dassonneville, Adèle Huguet, Marie-Laure Massot, Agnès Tricoche, Marie Carlin, Jean-Philippe Moreux, and Rostaing Aurélie. ‘Compte-rendu de la journée d’étude «Point HTR 2022» Transkribus / eScriptorium: Transcrire, annoter et éditer numériquement des documents d’archives’. Report, 9 June 2022. <https://hal.science/hal-03692413>
- [10] Huff, Dorothee, and Kristina Stöbener. ‘Projekt OCR-BW: Automatische Texterkennung von Handschriften’. o-bib. Das offene Bibliotheksjournal / Herausgeber VDB 9, no. 4 (29 November 2022): 1–19. <https://doi.org/10.5282/o-bib/5885>
- [11] Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, e Zhaopeng Tu. «Is ChatGPT A Good Translator? A Preliminary Study», 2023. <https://doi.org/10.48550/ARXIV.2301.08745>
- [12] Kahle, Philip, Sebastian Colutto, Gunter Hackl, e Gunter Muhlberger. «Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents». In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 19–24. Kyoto: IEEE, 2017. <https://doi.org/10.1109/ICDAR.2017.307>
- [13] Kaplan, Frédéric. «The Venice Time Machine». In Proceedings of the 2015 ACM Symposium on Document Engineering, 73–73. Lausanne Switzerland: ACM, 2015. <https://doi.org/10.1145/2682571.2797071>
- [14] Kaplan, Frédéric, e Isabella di Lenardo. «Big Data of the Past». Frontiers in Digital Humanities 4 (29 maggio 2017): 12. <https://doi.org/10.3389/fdigh.2017.00012>
- [15] Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. ‘EScriptorium: An Open Source Platform for Historical Document Analysis’. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 2:19–19, 2019. <https://doi.org/10.1109/ICDARW.2019.10032>
- [16] Lazzari, Gianrocco, Giovanni Colavizza, Fabio Bortoluzzi, Davide Drago, Andrea Erbooso, Francesca Zugno, Frédéric Kaplan, e Marcel Salathé. «A Digital Reconstruction of the 1630–1631 Large Plague Outbreak in Venice». Scientific Reports 10, fasc. 1 (20 ottobre 2020): 17849. <https://doi.org/10.1038/s41598-020-74775-6>
- [17] Maarand, Martin, Yngvil Beyer, Andre Kåsen, Knut T. Fosseide, and Christopher Kermorvant. ‘A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian’. In Document Analysis Systems, edited by Seiichi Uchida, Elisa Barney, and Véronique Eglin, 399–413. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022. [https://doi.org/10.1007/978-3-031-06555-2\\_27](https://doi.org/10.1007/978-3-031-06555-2_27)
- [18] Moretti, Giovanni, Rachele Sprugnoli, e Sara Tonelli. «Digging in the Dirt: Extracting Keyphrases from Texts with KD». In Proceedings of the Second Italian Conference on Computational Linguistics CLiC-It 2015, a cura di Cristina Bosco, Sara Tonelli, e Fabio Massimo Zanzotto, 198–203. Accademia University Press, 2015. <https://doi.org/10.4000/books.aaccademia.1518>
- [19] 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 (9 settembre 2019): 954–76. <https://doi.org/10.1108/JD-07-2018-0114>
- [20] Pavlik, John V. «Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education». Journalism & Mass Communication Educator 78, fasc. 1 (marzo 2023): 84–93. <https://doi.org/10.1177/10776958221149577>
- [21] Ponte, Mark, e Erik Schmitz. «Rembrandt Paints Master Carpenter Jacob Wesselsz Wiltingh». Kroniek van Het Rembrandthuis 2021, fasc. 1 (2021): 18–29. <https://doi.org/10.48296/KvhR2021.02>
- [22] Sánchez, Joan Andreu, Vicent Bosch, Verónica Romero, Katrien Depuydt, and Jesse de Does. ‘Handwritten Text Recognition for Historical Documents in the Transcriptorium Project’. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 111–17. DATeCH '14. New York, NY, USA: Association for Computing Machinery, 2014. <https://doi.org/10.1145/2595188.2595193>
- [23] Seaward, Louise, e Maria Kallio. «Transkribus: Handwritten Text Recognition Technology for Historical Documents», 2017. <https://dh-abstracts.library.cmu.edu/works/4193>
- [24] Sivasankari, M., e S. P. Victor. «Handwritten Text Recognition», 2014. <https://www.semanticscholar.org/paper/Handwritten-Text-Recognition-Sivasankari-Victor/513c052d1dd787367bf3343d3ad7ed2e4efdb7da>
- [25] Sobania, Dominik, Martin Briesch, Carol Hanna, e Justyna Petke. «An Analysis of the Automatic Bug Fixing Performance of ChatGPT», 2023. <https://doi.org/10.48550/ARXIV.2301.08653>

- [26] Terras, Melissa. «Inviting AI into the Archives: The Reception of Handwritten Recognition Technology into Historical Manuscript Transcription». In Archives, Access and Artificial Intelligence. Working with Born-Digital and Digitized Archival Collections, a cura di Lise Jaillant, 179–204. Bielefeld: Bielefeld University Press, 2022. <https://doi.org/10.1515/9783839455845-008>
- [27] Zhai, Xiaoming. «ChatGPT User Experience: Implications for Education». SSRN Electronic Journal, 2022. <https://doi.org/10.2139/ssrn.4312418>



# Codificare (al)l'Infinito

Beatrice Nava<sup>1</sup>, Roberta Priore<sup>2</sup>,

<sup>1</sup> Huygens Institute, Netherlands – beatrice.nava@huygens.knaw.nl

<sup>2</sup> Università di Bologna, Italia – roberta.priore2@unibo.it

## ABSTRACT

Il workshop si propone di affrontare, attraverso il caso di studio dell'*Infinito* di Giacomo Leopardi, il nodo centrale della marcatura dell'apparato nella costruzione di edizioni genetiche digitali, con l'obiettivo di stimolare la discussione e la condivisione di esperienze per stabilire un modello condiviso per la codifica di situazioni testuali simili. "Codificare (al)l'*Infinito*" rappresenta dunque sia un momento di riflessione teorica sui problemi testuali legati alla codifica di testi a tradizione pluritestimoniale (manoscritta e a stampa), sia un banco di prova pratico per la risoluzione delle criticità che tale pratica comporta. Verrà proposto infatti un modello di marcatura XML/TEI della poesia, con particolare attenzione alle problematiche riscontrate in fase di modellizzazione e alla descrizione delle diverse soluzioni alternative finora trovate, a cui seguirà un laboratorio di marcatura XML/TEI degli *Idilli* leopardiani che con l'*Infinito* fanno sistema all'interno del "quaderno napoletano" e, in fine, un esercizio di visualizzazione con *Edition Visualization Technology*. A questo lavoro, individuale o di gruppo ma guidato dalle proponenti, e dunque affrontabile a partire da livelli di partenza diversi, seguirà la messa in comune dei risultati, la condivisione di esperienze di codifica simili, di soluzioni emerse in corso d'opera o di eventuali problemi ulteriori, per la costruzione collettiva di soluzioni applicabili a più situazioni testuali che costituisce, come anticipato, lo scopo principale, e più ambizioso, della presente proposta.

## PAROLE CHIAVE

Edizione critica digitale; filologia d'autore; TEI, Leopardi

## 1. INTRODUZIONE

La diffusione crescente di progetti di edizioni scientifiche digitali e l'importanza centrale che tali prodotti hanno ormai assunto nella prassi quotidiana di ricerca e didattica e, in generale, come affidabile e facilmente accessibile fonte di informazione per un pubblico anche non specializzato, spingono sempre più studiosi, filologi e non, a progettare le loro edizioni non più o non solo in formato cartaceo, ma puntando sulle molteplici potenzialità che l'ambiente digitale è in grado di offrire. Ciò ha comportato, e il processo trasformativo è a nostro avviso ancora in pieno svolgimento, un vero e proprio cambio di paradigma editoriale, sia a livello metodologico e di modellizzazione<sup>1</sup> che di trattamento dell'informazione. Se infatti produrre un'edizione critica cartacea comporta già di per sé una qualche forma di codifica del contenuto da veicolare al lettore – si pensi ad esempio ai segni diacritici, ai simboli e alle abbreviazioni impiegati dall'editore negli apparati critici – la scelta di utilizzare il medium elettronico conferisce a tale prassi un'importanza ancor più centrale, essendo la modalità di codifica utilizzata per i propri dati la pietra angolare su cui poggia l'intera architettura dell'edizione, nonché il luogo in cui si originano le potenzialità computazionali della stessa.

Lo standard di codifica XML TEI, attualmente ineguagliato per completezza e flessibilità per la marcatura di informazioni testuali,<sup>2</sup> ha comportato, in particolare per la filologia d'autore, diversi problemi e questioni di difficile soluzione che riteniamo possano essere affrontati proficuamente solo entro un dibattito condiviso con tutte

---

<sup>1</sup> Si pensi, giusto per citare un esempio, al mutato rapporto rispetto ai materiali di partenza: la possibilità di rendere disponibili immagini digitali ad alta risoluzione e la disponibilità illimitata di spazio garantita dal supporto digitale generano immancabilmente riflessioni sulla necessità o meno di edizioni puramente documentarie-diplomatiche e insieme aprono alla possibilità/tentazione di includere nelle edizioni molteplici materiali o testimoni tradizionalmente espunti in sede di edizioni cartacee.

<sup>2</sup> Certo le critiche al vocabolario non mancano e non sono tutte ingiustificate. Tuttavia, ci pare che lo strumento sia, se mai, perfettibile più che da archiviare. Cfr. Cummings (2019).

e tutti coloro che si occupano o si interessano della questione, attraverso un approccio seminariale dal taglio pratico e in un'ottica collaborativa.<sup>3</sup>

Al momento, infatti, un confronto diretto con la marcatura TEI, nonostante le criticità, ci sembra in molti casi preferibile ad altre soluzioni, come ad esempio la codifica automatica mediante Domain Specific Languages (Bambaci, Boschetti 2020), che sfrutta giustamente l'alto grado di formalizzazione degli apparati critici – pure non così pacificamente estendibile alle edizioni di filologia d'autore, spesso ancora molto difformi sotto questo profilo – ma prevede una netta divisione tra i compiti del filologo tradizionale e quelli dello sviluppatore e del filologo digitale. Se da un lato tale approccio ha il vantaggio di semplificare le operazioni di codifica e di non richiedere competenze tecniche al filologo tradizionale, è altrettanto vero che un numero crescente di studiosi ha, o desidera sviluppare, «una vera e propria doppia competenza» (Leclerc, Saidi 2019, 144) e diversificare le competenze filologiche da quelle digitali, almeno a un livello basilare, può non essere sempre necessario e rischia forse di tenere il filologo in una posizione di non pieno controllo al momento cruciale della modellizzazione dei propri dati.

Per queste ragioni un workshop AIUCD – aperto in particolare a chi abbia affrontato o intenda affrontare i problemi di codifica di correzioni d'autore, ma anche in generale a chi si interessi di codifica di apparati critici in XML TEI – ci pare il luogo ideale per far nascere nuove idee e soluzioni condivise, a partire dal lavoro sui testi e dalla proposta del modello di codifica presentato. Per uno sguardo oltre la codifica, si concluderà il seminario con un cenno alle possibilità di visualizzazione offerte dai software open source, in particolare utilizzando EVT2 beta 2 (Rosselli Del Turco, Di Pietro 2019)<sup>4</sup> e discutendo delle potenzialità di EVT3 (a oggi non ancora rilasciato), pensato più nello specifico proprio per edizioni critiche di filologia d'autore (Rosselli Del Turco 2022).

## 2. OBIETTIVI E TARGET

Il workshop si pone due principali obiettivi, diretti a due possibili macrotipologie di partecipanti: a coloro che abbiano intenzione di preparare un'edizione critica digitale di filologia d'autore, ma non si siano direttamente ancora cimentati con la codifica, il workshop offre un primo campo di sperimentazione pratica e di riflessione metodologica, consentendo sia di apprendere i rudimenti della codifica XML TEI per la rappresentazione di correzioni d'autore e varianti a stampa, sia di riflettere più in generale sui possibili problemi dell'opera a cui hanno intenzione di lavorare, che potranno essere presentati e discussi nella fase finale del seminario. Per coloro invece che si siano già cimentati in edizioni critiche di questo tipo, lo scopo è quello di arrivare, attraverso le loro esperienze e proposte e tramite la discussione del caso di studio presentato, a ipotizzare un modello di codifica il più possibile condiviso per la rappresentazione di un apparato genetico che si auspica possa contribuire alla formazione di un modello più generale per le edizioni critiche digitali di opere pluritestimoniali con correzioni d'autore, nel rispetto, naturalmente, delle singole specificità testuali.

## 3. ARTICOLAZIONE

Al centro del workshop è la presentazione di un caso di studio specifico, l'*Infinito* di Giacomo Leopardi, del quale verrà mostrata una possibile marcatura XML/TEI<sup>5</sup> con particolare attenzione alle problematiche riscontrate in fase di modellizzazione e alla descrizione delle diverse soluzioni alternative finora trovate. La scelta di un testo come l'*Infinito* ci pare strategica per la sua brevità, connessa però a una storia di studio e di edizioni molto importante, ma anche per la storia testuale, dal momento che la sua tradizione manoscritta coinvolge più testi, tutti gli *Idilli* (1819-1821) – composti su quello che conosciamo come 'quaderno napoletano' con un meccanismo peculiare di stesura in tre tempi (Italia 2016, pp. 147-184) – che costituiranno i materiali di partenza per la parte pratica del workshop.

Dopo una breve introduzione teorica alla storia del testo, all'apparato genetico e ai problemi di marcatura, si entrerà perciò nel vivo delle questioni di codifica XML/TEI per la filologia d'autore (Italia 2019), partendo dalla

---

<sup>3</sup> Non è forse un caso che la maggior parte delle edizioni digitali basate su una codifica XML/TEI presenti di fatto solo una trascrizione diplomatica, mentre più rare sono le edizioni critiche digitali e sostanzialmente assenti quelle di filologia genetica. Cfr. Rosselli Del Turco, Di Pietro (2019, 152) e Nava (2022).

<sup>4</sup> Cfr. anche Rosselli Del Turco, Di Pietro, Martignano (2019).

<sup>5</sup> Il modello è stato elaborato nel corso di alcuni incontri seminariali tenutisi presso l'Università di Bologna, cui hanno collaborato anche Ilaria Burattini, Giulia Menna, Filippo Pelacci e David Tagliacozzo.

presentazione della marcatura dell'opera campione, con accento specifico sui nodi più spinosi tipici della rappresentazione in XML/TEI di materiali di questo tipo, come i casi di *overlapping*, il connesso problema del 'riuso' – ossia di quelle situazioni in cui l'autore conserva parte di una lezione scartata nella riformulazione successiva – e, più in generale, la difficoltà di casi specifici in cui le correzioni si infittiscono e la marcatura, di conseguenza, pare moltiplicarsi all'infinito.

Seguirà poi una fase di laboratorio vera e propria durante la quale sarà chiesto alle e ai partecipanti di marcare, secondo il modello di codifica proposto, uno degli altri componimenti racchiusi nel 'quaderno'. Dopo aver messo il modello alla prova di più testi, tra loro affini, si aprirà la fase di dibattito e restituzione finale, in cui potranno essere condivise esperienze di codifica simili, soluzioni emerse in corso d'opera o problemi ulteriori, per la costruzione collettiva di soluzioni di codifica condivise e applicabili a più situazioni testuali che costituisce, come detto, lo scopo principale, e più ambizioso, della presente proposta.

La volontà è quella di tentare di superare, attraverso il raggiungimento di queste soluzioni condivise (o anche solo grazie al ponte tra studiosi e studiosi che questo workshop potrà gettare per discussioni future) il costante statuto di "work in progress" dei progetti di edizioni digitali, nei quali il processo di definizione del modello di marcatura, sempre potenzialmente discutibile, rischia di rivelarsi scoraggiante.

A livello più immediatamente pratico, l'articolazione del seminario sarà la seguente: 15' di introduzione sullo stato dell'arte delle edizioni digitali di filologia d'autore e sui principali problemi e casi virtuosi riscontrabili; 15' di presentazione del caso del "quaderno napoletano" e 15' di descrizione della proposta di marcatura dell'*Infinito*. Seguirà poi la fase laboratoriale di 45' di marcatura degli altri testi del "quaderno" e visualizzazione con EVT (con supporto diretto ed eventuali spiegazioni introduttive per chi fosse alle prime armi in tal senso) e, in fine, mezz'ora di discussione finale dei problemi riscontrati in fase di codifica e condivisione di proposte modellizzanti.

#### 4. PREREQUISITI E STRUMENTI

Si richiede una conoscenza, anche base, del linguaggio di codifica XML/TEI e un editor xml (sublime, oxygen, ecc.). Coloro che fossero interessati a discutere eventuali situazioni testuali particolari, nell'ultima fase del workshop, potranno presentare i loro materiali di lavoro.

#### BIBLIOGRAFIA

- [1] Bambaci, Luigi e Federico, Boschetti. Encoding critical Encoding the Critical Apparatus by Domain Specific Languages: The Case of the Hebrew Book of Qohelet. Quaderni di Umanistica digitale (2020): 7-13.
- [2] Cummings, James. A world of difference: Myths and misconceptions about the TEI. Digital Scholarship in the Humanities 34 (2019): 158-179 <https://doi.org/10.1093/llc/fqy071>
- [3] Italia, Paola. Il metodo di Leopardi. Carocci, 2016.
- [4] Italia, Paola, Filologia d'autore digitale. Ecdotica 16 (2019): 202-216.
- [5] Leclerc, Élise e Saïdi, Samantha. Storia di un ornitorinco: note sulla collaborazione in progetti di edizione scientifica digitale. Ecdotica 16 (2019): 202-216.
- [6] Nava, Beatrice. Siamo tutti bédieriani? Prospettive per le edizioni genetiche digitali. Umanistica Digitale 6,14 (2022): 19-40. <https://doi.org/10.6092/issn.2532-8816/14949>
- [7] Rosselli Del Turco Roberto. EVT - Edition Visualization Technology 3 (v. 1.0.0-alpha), GitHub, <https://github.com/evt-project/evt-viewer-angular>
- [8] Rosselli Del Turco, Roberto Di Pietro, Chiara. La visualizzazione di edizioni digitali con EVT. Ecdotica 16 (2019): 134-147.
- [8] Rosselli Del Turco, Roberto, Di Pietro, Chiara e Martignano, Chiara. Progettazione e implementazione di nuove funzionalità per EVT 2: lo stato attuale dello sviluppo. Umanistica Digitale 3, 7 (2019): 5-21. <https://doi.org/10.6092/issn.2532-8816/9322>

ASSOCIAZIONE per  
l'INFORMATICA UMANISTICA  
e la CULTURA DIGITALE



UNIVERSITÀ  
DI SIENA 1240



SANTA CHIARA LAB  
Università di Siena 1240

CENTRO DI STUDI COMPARATI  
IDEUG-SU



Centro di Ricerca  
Franco Fortini



DISPOC

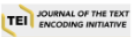


DIISM



DSSBC

Con il patrocinio di



JOURNAL OF THE TEXT  
ENCODING INITIATIVE